MOLECULAR ANALYSES OF AVIAN SEX CHROMOSOMES AND SEX CHROMOSOME-LIKE AUTOSOMES

A Dissertation
Presented to
The Academic Faculty

by

Dan Sun

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy in Bioinformatics

Georgia Institute of Technology

August 2019

MOLECULAR ANALYSES OF AVIAN SEX CHROMOSOMES AND SEX CHROMOSOME-LIKE AUTOSOMES

Approved by:

Dr. Soojin V. Yi, Advisor
School of Biological Sciences
Georgia Institute of Technology

Dr. Gregory C. Gibson
School of Biological Sciences
Georgia Institute of Technology

Dr. I. King Jordan
School of Biological Sciences
Georgia Institute of Technology

Dr. Donna L. Maney
Department of Psychology
Emory University

Dr. Joseph Lachance
School of Biological Sciences
Georgia Institute of Technology

Date Approved: July 26, 2019

To girls who are not afraid to pursue their dreams..

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Soojin Yi. I am deeply indebted to her invaluable input, advice, and support; and the way how she strives for excellence in research is highly motivational and inspirational. I also want to thank my thesis committee members Dr. Donna Maney, Dr. Joseph Lachance, Dr. King Jordan, and Dr. Greg Gibson for reading my thesis and providing very useful suggestions and corrections.

I am truly grateful to my wonderful colleagues from the Yi Lab. I would like to thank Dr. Isabel Mendizabal, Dr. Iksoo Huh, Dr. Jia Zeng, Dr. Ke Xu, Paramita Chatterjee, Thomas Layman, Hyeonsoo Jeong, Xin Wu, Devika Singh, and Dr. Thomas Keller for encouragement, research insights, and career advice.

I would like to express my deepest appreciation to my family and friends. Throughout the seven years as a graduate student at Georgia Tech, I visited my family in China only four times. Their encouraging long-distance text messages and phone calls got me through the toughest time of my graduate life. I was lucky enough to have met my husband Chen Feng at Georgia Tech. He is the most patient and rational person I know and has taught me so much. My thanks also go out to all my friends who have brought me so much happiness.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

5mC	5-methylcytosine
d_C	Conservative amino acid substitution rate
d_N	Non-synonymous substitution rate
d_P	Promoter substitution rate
d_R	Radical amino acid substitution rate
d_S	Synonymous substitution rate
d_{XY}	Genetic divergence
ASE	Allele-specific gene expression
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
CARTPT	Cocaine- and amphetamine-regulated transcript
cDNA	complementary DNA
DMC	Differentially methylated CpG
DMRT1	Doublesex and mab-3 related transcription factor 1
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
E	Embryonic day
F	Female
FDR	False discovery rate
FISH	Fluorescent in situ hybridization
F_{ST}	Fixation index
GO	Gene ontology
Hyper	Hypermethylated
Hypo	Hypomethylated
IQR	Interquartile range

lncRNA	Long non-coding RNA
M	Male
MHM	Male hypermethylated
NS	Not significant
PC	Principal component
PCA	Principal component analysis
PE	Paired-end
RBP	RNA binding protein
RLN3	Relaxin 3
RNA-seq	RNA sequencing
S	Stratum
SNP	Single nucleotide polymorphism
TE	Transposable element
TES	Transcription end site
TGU	*Taeniopygia guttata*
TPM	Transcript per kilobase million
TSS	Transcription start site
VLDLR	Very low-density lipoprotein receptor
WGBS	Whole-genome bisulfite sequencing
WGS	Whole-genome sequencing
ZAL	*Zonotrichia albicollis*

SUMMARY

Sex chromosomes have originated multiple times throughout eukaryotes. In species with the XY sex-determination system, dosage compensation (a process that balances expression of sex-linked genes between sexes) is often efficient, and its epigenetic basis has been well studied. However, the extent of epigenetic differentiation between sexes in female-heterogametic systems (ZW), which generally lack complete compensation, is poorly understood. Here, I examined the genome-wide DNA methylation landscapes between males and females in mammalian and avian species. In contrast to the X chromosome in mammals, birds display highly similar methylation patterns between sexes on the Z chromosome. Despite this, in chicken and potentially other species in the Galloanserae lineage, two extremely localized regions with pronounced methylation differentiation were observed, including a previously identified locus (referred to as 'male hypermethylated [MHM1]') and a novel locus (referred to as 'MHM2'). The two MHM loci bear remarkably similar molecular features and potential function in reducing male-to-female expression ratios of their neighboring genes. Therefore, DNA methylation is employed to solve dose problems for genes potentially essential to females, at least twice in the evolutionary history of the Galloanserae lineage.

In the white-throated sparrow, a pair of autosomes that are distinguished by chromosomal inversions resemble sex chromosomes. In this species, two plumage morphs that mate almost exclusively with each other display striking behavioral differences: within the same sex, birds of the white-striped morph (ZAL2/ZAL2^m) display more territorial aggression and less nestling provision than birds of the tan-striped morph (ZAL2/ZAL2). A detailed genomic comparison between a tan bird and a rare ZAL2^m homozygote revealed subtle nucleotide differences between ZAL2 and ZAL2^m as well as

weak degeneration of the non-recombining ZAL2^{m} chromosome. Nevertheless, a large proportion of genes exhibit allelic differential expression in the brain. Intriguingly, similar to the evolutionary path taken by sex chromosomes across many taxa, dosage compensation evolved as a mechanism to re-balance expression between morphs in this nascent autosomal system.

Last, I examined the DNA methylation landscape of the white-throated sparrow. Differences in DNA methylation between chicks and adults are pervasive across the genome, with hypermethylation in adults consistent with the overexpression of DNA methyltransferases. Functional enrichment analysis revealed that the observed changes in methylation are likely involved in development. In contrast to the widespread age effects, morph influences are most prominent on the ZAL2/ZAL2^{m} chromosomes. Notably, allelic differences in DNA methylation and allelic differences in gene expression are significantly linked. Taken together, these findings offer new insights into the epigenetic regulation of gene expression in avian sex chromosomes and sex chromosome-like autosomes.

CHAPTER 1

INTRODUCTION

Sex chromosomes have evolved independently multiple times across eukaryotes (Bachtrog et al. 2014; Wright et al. 2016). Unlike the male-heterogametic system in mammals (males: XY; females: XX), birds have a female-heterogametic system (males: ZZ; females: ZW) (Zhou et al. 2014). Avian Z and W chromosomes evolved from autosomes ~140 million years ago (Mya) by a series of chromosomal rearrangements (Nam and Ellegren 2008; Cortez et al. 2014). These events were accompanied by a stepwise cessation of recombination between Z and W, leading to the degeneration of the non-recombining W chromosome. In chicken, for example, only 28 out of 685 genes (~4%) survived genetic decay (Bellott et al. 2017). Haploinsufficiency due to the degeneration of the non-recombining sex chromosome in the heterogametic sex may trigger the evolution of dosage compensation. Indeed, the reduced dosage is often compensated in mammals or other species with an XY system (Mank 2013; Mullon et al. 2015). The ZW system has arguably received much less attention from the researchers. Nevertheless, studies so far indicate that dosage compensation is incomplete in birds (Ellegren et al. 2007; Itoh et al. 2007; Graves 2016).

In the well-studied mammalian system, epigenetic mechanisms such as DNA methylation are causatively implicated in dosage compensation (Sharp et al. 2011; Cotton et al. 2015; Graves 2016). DNA methylation is an epigenetic mark involving the addition of a methyl group to the fifth carbon of the cytosine, which is facilitated by a group of enzymes named DNA methyltransferases (DNMTs) (Bird 1992; Bird 2002). Among these DNMTs, DNMT1 is engaged in the *de novo* addition of methyl groups, whereas DNMT3a and DNMT3b play maintenance roles by adding methyl groups to hemimethylated DNA

strands after replication (Okano et al. 1999; Hermann et al. 2004). DNA methylation is involved in numerous processes such as epigenetic silencing of transcription, X inactivation, genomic imprinting, suppression of transposable elements, development and aging, and sex differentiation (Robertson and A.Jones 2000; Roeszler et al. 2012; Sun and Yi 2015; Graves 2016). Disruption of DNA methylation is implicated in various cancer types and developmental disorders (Robertson 2005; Hamidi et al. 2015).

In species with the XY sex-determination system, the epigenetic basis for dosage compensation is well understood. However, the extent of epigenetic differentiation between the sex chromosomes in female-heterogametic systems, which typically lacks complete dosage compensation, is relatively understudied. For instance, the only known region of epigenetic differentiation between chicken sex chromosome is a locus termed 'male hypermethylated (MHM),' which was identified using cDNA clones and methylation-sensitive restriction enzymes as a region of extremely higher methylation in males than in females on the Z chromosome (Teranishi et al. 2001). Consistent with its methylation status, the MHM locus is transcriptionally inactive in males but expresses high-molecular-weight long non-coding RNAs (lncRNAs) in females in a non-tissue-specific fashion. Therefore, the observation of reduced male-to-female expression ratios near the MHM locus leads to the conjecture that MHM lncRNAs may bind to surrounding genes to increase their expression in females (Melamed and Arnold 2007). However, these hypotheses were not fully explored due to the lack of studies on the comprehensive genome-wide characterization of epigenetic divergence between the sex chromosomes. In Chapter 2, we take advantage of published epigenetic data, including whole-genome bisulfite sequencing (Uebbing et al. 2015; Laine et al. 2016; Lee et al. 2017) and assay for transposase accessible chromatin using sequencing (ATAC-seq) data (Foissac et al. 2018; Sackton et al. 2018), to investigate the role of epigenetic divergence on sex

chromosome dosage in chicken. In addition, we employ large-scale RNA sequencing (RNA-seq) data from multiple tissues across several developmental stages (more than 150 samples) (Brawand et al. 2011; Julien et al. 2012; Ayers et al. 2013; Uebbing et al. 2015; Zimmer et al. 2016; Marin et al. 2017) to understand the sex-specific expression patterns across the Z chromosome.

The third and fourth chapters include analyses of a newly evolving sex chromosome-like autosomal pairs in the white-throated sparrow (*Zonotrichia albicollis*). In this species, two plumage morphs occur with striking behavioral differences: within the same sex, birds of the white-striped morph (hereafter "white") display more territorial aggression and less nestling provision than birds of the tan-striped morph (hereafter "tan") (Tuttle 2003; Maney 2008; Horton et al. 2014a; Horton et al. 2014b; Maney et al. 2015; Zinzow-Kramer et al. 2015). The phenotypic dimorphism can be traced to a pair of chromosomal inversions harboring ~1,000 genes, that occurred approximately 2-3 million years ago (Thorneycroft 1975; Thomas et al. 2008; Huynh et al. 2010). Birds of the white morph are heterozygous for the rearrangement (genotype: ZAL2/ZAL2^{m}; 'm' stands for metacentric), and birds of the tan morph are homozygous for the standard karyotype (genotype: ZAL2/ZAL2).

Interestingly, white and tan birds mate in a strongly disassortative manner, leading to the suppressed recombination of ZAL2^{m} (Thomas et al. 2008; Davis et al. 2011; Huynh et al. 2011). Thus, the ZAL2 and ZAL2^{m} system resembles the sex chromosomes in that one chromosome (ZAL2^{m}) is transmitted through only approximately half of the total population and that it largely lacks recombination. Indeed, a recent study revealed the degeneration of the ZAL2^{m} chromosome (Tuttle et al. 2016). Nevertheless, the ZAL2^{m} and ZAL2 chromosomes are similar in their sizes, and homozygotes for ZAL2^{m} exist in nature (Thorneycroft 1975; Horton et al. 2013; Tuttle et al. 2016). For this incipient system in

which recombination is reduced but not entirely lacking, we aim to understand: 1) the degree of genetic degeneration, 2) the degree of expression divergence between alleles, 3) whether dosage compensation, a feature shared by sex chromosomes of diverse taxa, has evolved in response to disrupted gene expression, and 4) whether epigenetic mechanisms such as DNA methylation play a role in the regulatory evolution of the ZAL2/ZAL2^{m} system.

As birds from the white-striped morph are typically heterozygous for the second chromosome, directly comparing ZAL2 and ZAL2^{m} sequences has been challenging. In Chapter 3, by taking advantage of whole-genome sequences of a rare super-white individual homozygous for the ZAL2^{m} chromosome (occurs at a ~0.2% frequency in nature) (Horton et al. 2013) and a tan reference genome (ZAL2/ZAL2) (Tuttle et al. 2016), we compare ZAL2 and ZAL2^{m} at the chromosome-wide scale to characterize the extent of ZAL2^{m} degeneration. Meanwhile, Zinzow-Kramer et al. (Zinzow-Kramer et al. 2015) previously generated and analyzed brain RNA-seq data from both morphs to identify weighted co-expression networks linked to singing behavior at the morph level. Using this dataset, we identify allele-specific expression patterns and seek to answer the question of whether dosage compensation exists in response to potentially disrupted ZAL2^{m} expression.

DNA methylation has long been considered a fundamental epigenetic mechanism for gene regulation (Bird 1992; Bird 2002). In Chapter 4, by utilizing newly sequenced WGBS data of 12 sparrows from the two morphs, we aim to understand whether DNA methylation drives allele-specific expression. In addition, we explore how other factors such as age could interweave with morph to shape the methylation landscape in the white-throated sparrow. In summary, our comprehensive genetic, epigenetic, and transcriptomic

studies will shed light on the regulatory evolution of avian sex chromosomes and sex chromosome-like autosomes.

CHAPTER 2

REGIONAL EPIGENETIC DIFFERENTIATION OF THE CHICKEN Z CHROMOSOME

2.1 Introduction

Chromosomal sex determination has originated multiple times in different lineages of animals (Bachtrog et al. 2014; Graves 2016). For example, mammals have a male-heterogametic system (females: XX; males: XY), while birds have a female-heterogametic system (males: ZZ; females: ZW) (Zhou et al. 2014; Graves 2016). The common path of the evolution of most sex chromosome systems is a step-wise cessation of recombination between the proto-sex chromosomes (Charlesworth 1978; Lahn and Page 1999), leading to the distinct evolutionary strata of the non-recombining chromosome (Lahn and Page 1999; Nam and Ellegren 2008; Zhou et al. 2014). The reduced recombination is accompanied by declined effective population size of the non-recombining chromosome, causing genetic degeneration (Charlesworth and Charlesworth 2000; Yi and Charlesworth 2000; Bachtrog 2013).

Following degeneration, haploinsufficiency in the heterogametic sex may trigger the evolution of dosage compensation. In male-heterogametic systems, dosage compensation is often global and nearly complete. For instance, the X chromosome is largely inactivated in females in eutherian mammals (Carrel and Willard 2005; Graves 2016). These alterations are accompanied by epigenetic mechanisms such as DNA methylation and histone modifications (Hellman and Chess 2007; Brockdorff and Turner 2015; Lucchesi and Kuroda 2015; Marin et al. 2017).

Although reduced dosage is often compensated in mammals or other species with an XY system, dosage compensation is incomplete in birds (Ellegren et al. 2007; Itoh et al. 2007; Adolfsson and Ellegren 2013; Wang et al. 2014; Mullon et al. 2015; Graves 2016; Marin et al. 2017). Z-linked genes in general have higher expression in males (with two Z chromosomes) than in females (with one Z chromosome), with only a subset of genes that are dosage-sensitive exhibiting similar expression levels between sexes (Itoh et al. 2007; Itoh et al. 2010; Uebbing et al. 2013; Uebbing et al. 2015). Consequently, the sex-specific epigenetic profile of Z chromosomes in female-heterogametic systems is likely to be different from that of the X chromosomes in male-heterogametic systems.

Despite the lack of global dosage compensation on the Z chromosome, chicken microarray data and subsequent RNA-seq analysis in the Galloanserae lineage (chicken and their close relatives) have revealed a localized region surrounding the male hypermethylated (MHM) locus on the Z chromosome, which displays a pronounced reduction in male-to-female expression ratios (Melamed and Arnold 2007; Mank and Ellegren 2009; Melamed et al. 2009; Wright et al. 2015). The MHM locus was previously recognized using a targeted approach and is composed of large tandem arrays on the Z chromosome in Galloanserae species (Teranishi et al. 2001). Consistent with its methylation status, the MHM locus is transcriptionally inactive in males but expresses high-molecular-weight long non-coding RNAs (lncRNAs) in females in a non-tissue-specific fashion. Therefore, the observation of reduced male-to-female expression ratios near the MHM locus leads to the conjecture that MHM lncRNAs may bind to the surrounding genes to increase their expression of in females (Melamed and Arnold 2007).

The global epigenetic characterization of the sex chromosomes in avian species has been impeded by the late emergence of their whole-genome methylomes (Li et al. 2015; Mugal et al. 2015; Laine et al. 2016; Lee et al. 2017; Zhang et al. 2017), and sex

chromosomes were neglected in all of these studies. In this study, to understand the extent of epigenetic divergence between male and female Z chromosomes, we investigated large-scale whole-genome methylation (18 samples) and chromatin accessibility maps (32 samples) as well as transcriptome data (181 samples) from chickens and other outgroup species (Table A. 1). We will delve into the following questions: 1) Is DNA methylation a mechanism to mediate W degeneration? 2) Based on the distinct degrees of dosage compensation between female- and male-heterogametic sex chromosomes, how are the epigenetic profiles different? 3) Does MHM represent the only epigenetic locus of differentiation between sexes in chicken and other Galloanserae species? 4) Do the localized sex differences in methylation also occur in other avian lineages, and hence potentially a common feature of species with female-heterogametic sex chromosomes? 5) Did MHMs originate by male-hypermethylation or female-hypomethylation?

2.2 Results

2.2.1 DNA methylation and gene expression in avian genomes

First, we characterized the relationship between DNA methylation and gene expression in avian genomes. We used brain whole-genome DNA methylation and transcriptome data from the same developmental stage in chicken (embryonic day 18 [E18]) (Uebbing et al. 2015; Lee et al. 2017) and a male great tit adult (Laine et al. 2016). CpG methylation and gene expression display negative correlations not only at promoters but also at gene bodies (Figure 2.1A). This observation supports the idea that the model

of transcriptional silencing by DNA methylation (Schübeler 2015) can be applied to most loci in avian genomes.

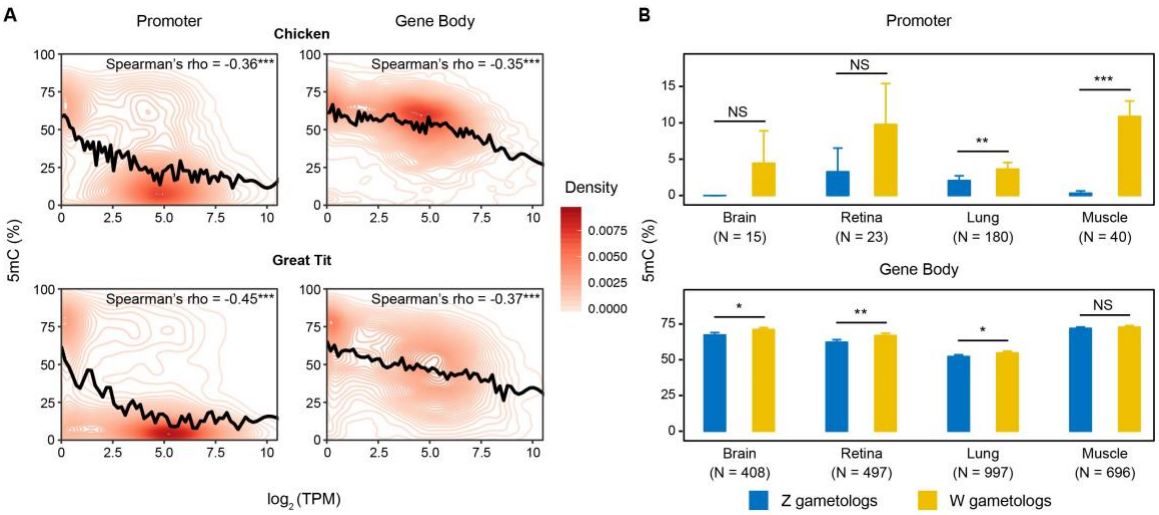

Figure 2.1 DNA methylation and expression in avian genomes. (A) Density of data points as a function of CpG methylation (5mC [%]) and gene expression (log_2 [TPM]) in chicken and great tit samples. The relationship between methylation and gene expression was smoothed with cubic splines (black lines). Number of genes with CpG methylation data: chicken: N = 11,662 for promoters and N = 11,723 for gene bodies; great tit: N = 14,694 for promoters and N = 14,721 for gene bodies. **(B)** Comparison of DNA methylation between aligned CpGs of Z and W gametologs. N depicts the number of Z-W-aligned CpGs with at least three mapped reads in each sample. Statistical significance was evaluated using paired Mann–Whitney *U* tests. For (A) and (B), the promoter of a gene was defined as -1.5 kb ~ +500 bp of its transcription start site. ***: $P < 0.001$; **: $P < 0.01$; *: $P < 0.05$; NS: not significant.

Next, based upon this relationship, we asked whether DNA methylation was associated with differential expression of the Z and W gametologs (homologous genes on the sex chromosomes). In the surviving pairs ($N = 26$, Methods), both promoters and gene bodies were significantly more heavily methylated on the W chromosome than the Z chromosome (Figure 2.1B), concordant with lower expression of W relative to Z gametologs (Figure 2.2). This pattern was consistent across tissue types and

developmental stages (Figure 2.1B), although the differences were not significant in some samples (brain and retina E18) which were likely due to low sequencing depths. This observation indicates that DNA methylation may act as a mechanism to silence W gametologs.

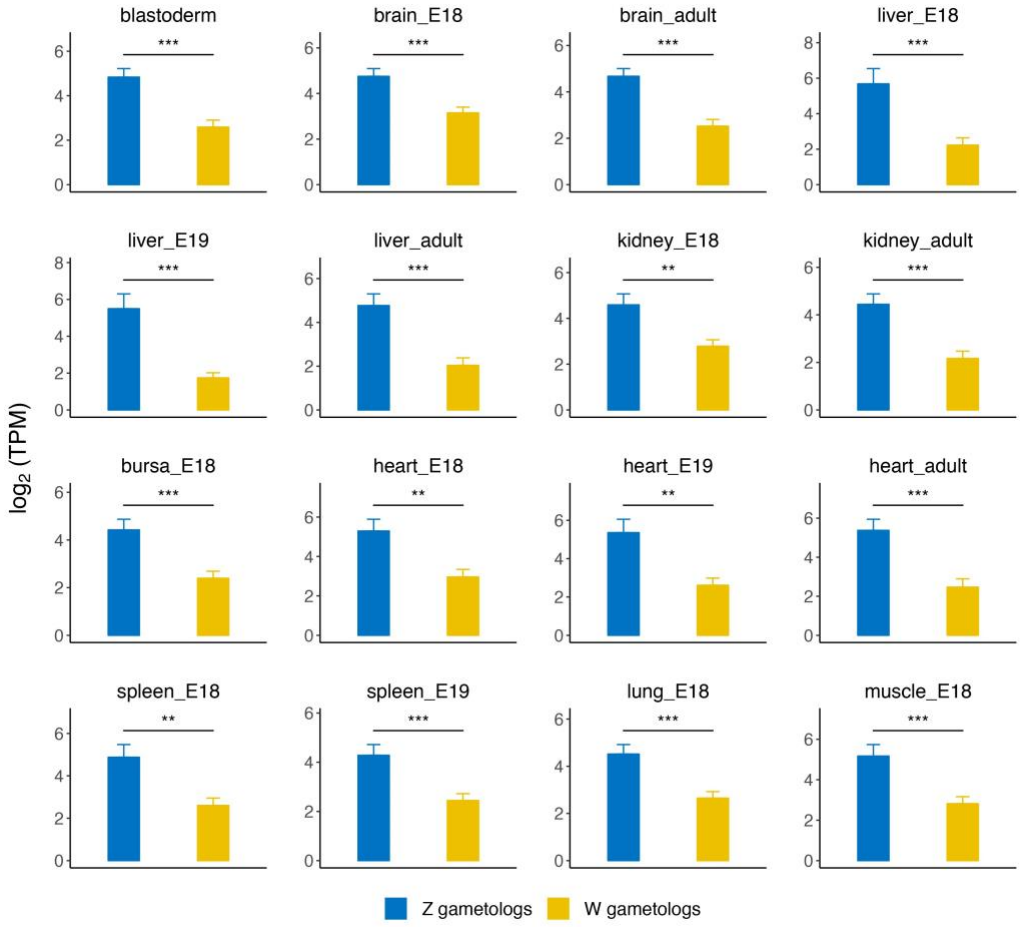

Figure 2.2 Comparison of expression levels of Z-W pairs. Z gametologs exhibit consistently higher expression than W gametologs. Statistical significance was evaluated using Mann–Whitney U tests (***: $P < 0.001$; **: $P < 0.01$).

2.2.2 A novel locus exhibiting striking DNA methylation and chromatin accessibility difference between male and female chicken Z chromosomes

Do DNA methylation profiles between sexes differ between female- and male-heterogametic sex chromosomes and between different avian lineages? To answer these questions, we investigated whole-genome bisulfite sequencing (WGBS) data of the human, chicken (Galloanserae), and white-throated sparrow (Neoaves). In contrast to the pattern observed in the human X chromosome (Figure 2.3A), DNA methylation between male and female Z chromosomes were highly similar to each other in both avian lineages (Figure 2.3B-C).

Despite this commonality, in chicken, but not the sparrow, we detected two notable exceptions with male-biased methylation patterns consistently across tissues (Figure 2.3B-C; Figure 2.4; Methods). The first region localizes to 27.140 Mb – 27.398 Mb of the chicken Z chromosome (Figure 2.3B-D; Figure 2.4). This region corresponded to a previously identified region using cDNA clones and methylation-sensitive restriction enzymes and was referred to as 'male-hypermethylated locus (MHM)' (Teranishi et al. 2001; Itoh et al. 2010). We discovered another region on the Z chromosome with striking sex differences in methylation across all tissue types examined (Figure 2.3B-D; Figure 2.4). This novel region was located in 73.160 Mb – 73.173 Mb of the Z chromosome. We will refer to the previously identified MHM as 'MHM1' and this newly identified region as 'MHM2'. Additionally, we examined recent ATAC-seq data (Foissac et al. 2018; Sackton et al. 2018) and found that at both MHM1 and MHM2, chromatin accessibility was elevated in females (Figure 2.3E**;** Figure 2.5).

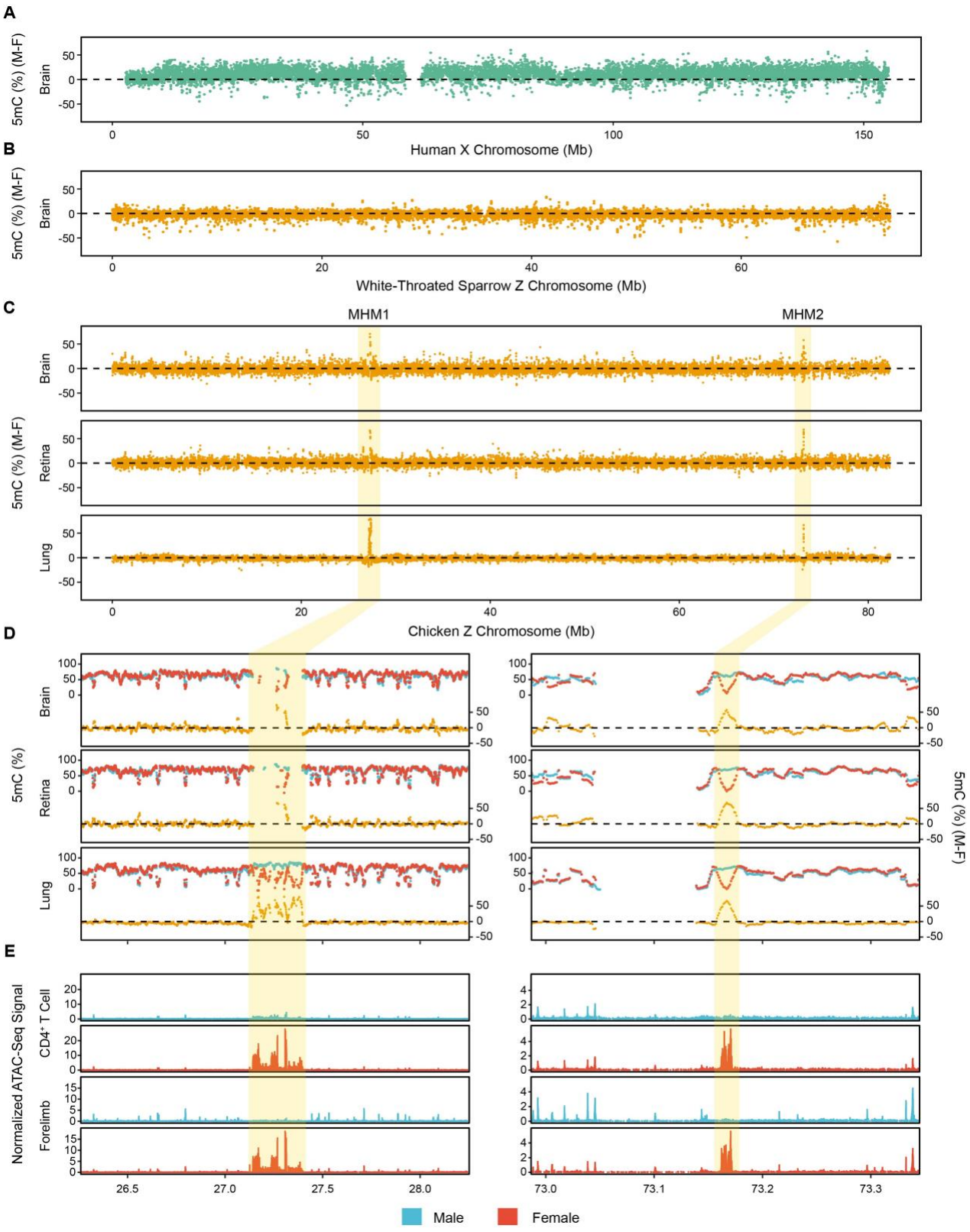
A
5mC (%) (M-F)
Brain
Human X Chromosome (Mb)
B
White-Throated Sparrow Z Chromosome (Mb)
C
MHM1
MHM2
Retina
Lung
Chicken Z Chromosome (Mb)
D
5mC (%)
E
Normalized ATAC-Seq Signal
CD4+ T Cell
Forelimb
Male
Female

Figure 2.3 Cross-lineage DNA methylation maps of X or Z chromosomes and chromatin accessibility patterns of MHMs. (A) Differences in DNA methylation between male and female X chromosomes in human brains (5mC [%] [M-F]) using data from (Zeng et al. 2012). **(B)** Differences in DNA methylation between male and female Z chromosomes in the white-throated sparrow. **(C)** Differences in DNA methylation between male and female Z chromosomes in chicken. Two outlier regions (MHM1 and MHM2, see Methods for identification) are highlighted. For A-C, methylation values were plotted using a 10 kb window size with a 1 kb step size. **(D)** A zoomed-in view of the MHM loci. Methylation levels (5mC [%]) for males (blue) and females (red) are shown in the upper lines. Lower lines (orange) depict sex differences in DNA methylation. **(E)** Both MHMs (shaded areas) display increased chromatin accessibility in females. ATAC-seq reads were merged per sex and normalized to fragment pileup per million reads for direct comparison between sexes. For either $CD4^+$ T cells (Foissac et al. 2018) or forelimb (E4.5) (Sackton et al. 2018), the two loci contain significantly female-biased peaks, tested using bdgdiff from the MACS2 program (Zhang et al. 2008). The vast majority of tissue/cell types with available data show similar patterns (Figure 2.5).

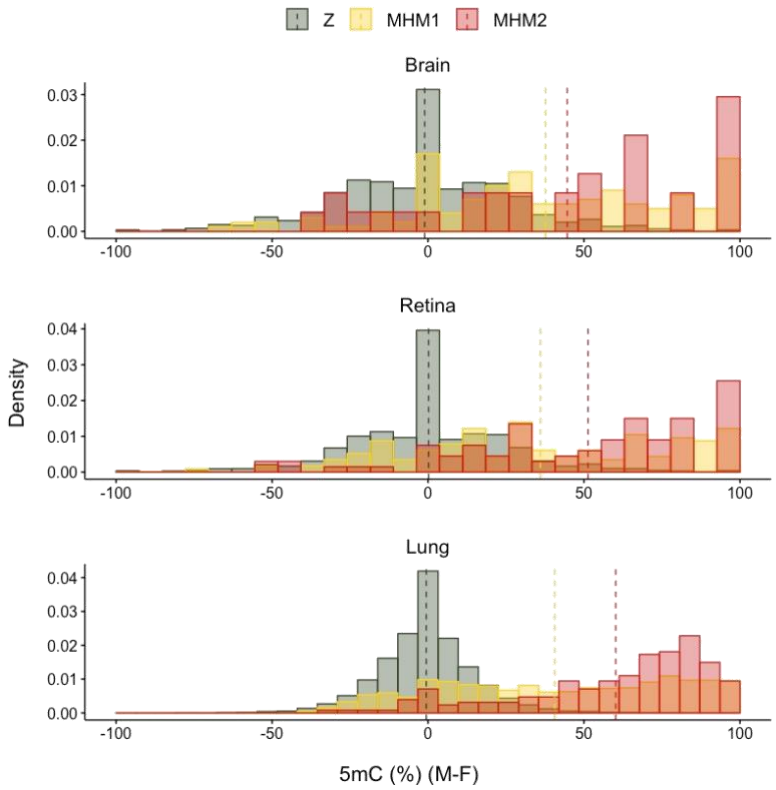

Figure 2.4 Two loci exhibit higher DNA methylation in males than in females on the chicken Z chromosome. Dashed lines depict the mean differences between males and females. Although the average differences in DNA methylation between males and females for the Z chromosome are close to zero, both MHMs are strongly male-hypermethylated.

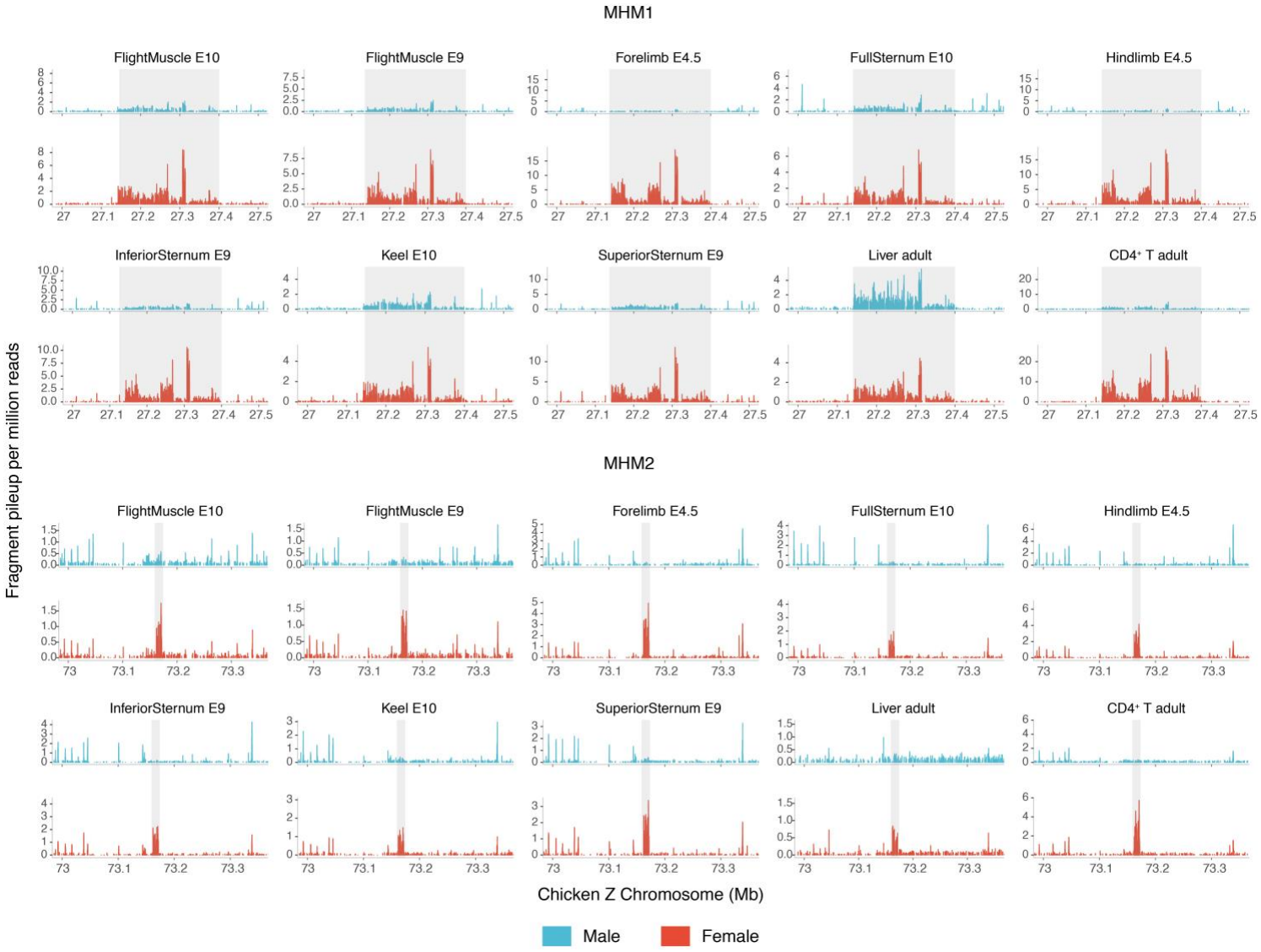

Figure 2.5 Both MHM loci (shaded areas) display higher chromatin accessibility in females compared with males across multiple tissues/cell types. ATAC-seq reads were merged per sex and normalized to fragment pileup per million reads for direct comparisons between sexes. Except for the MHM1 locus of adult liver samples, both MHM loci contain significantly female-biased peaks across tissues/cell types, tested using bdgdiff from the MACS2 program (Zhang et al. 2008). The source for liver and CD4$^+$ T cell samples is (Foissac et al. 2018) and for other tissue samples is (Sackton et al. 2018).

2.2.3 Shared molecular characteristics between MHM1 and MHM2

Although the alignment between the MHM1 and MHM2 loci revealed no shared homology, which hinted at their independent origins, the novel MHM2 locus exhibited distinct characteristics that paralleled those of the MHM1. First, both MHM loci were highly repetitive with unique repeat structures. Specifically, the MHM1 locus was comprised of tandem arrays of several long repeat units (1.8 – 3 kb) across an extended genomic region (250 Mb), while the MHM2 locus was comprised of four repetitive blocks of tandem repeats, with three to four iterations of a 542 bp repeat unit per block (Figure 2.6A). Second, by *de novo* annotation of transcripts using RNA-seq data from 165 chicken samples across tissues and developmental stages, we found that both loci encoded long non-coding RNAs (lncRNAs) (Figure 2.7). Specifically, MHM1 transcripts (only poly[A]-transcripts for the limitation of library selection methods used by these datasets) were spread across the whole locus, but three out of five MHM2 transcripts were located within the second repetitive block. As expected from the patterns of epigenetic differences, the expression of MHM1 and MHM2 lncRNAs were both female-biased in the vast majority of samples (Figure 2.6B).

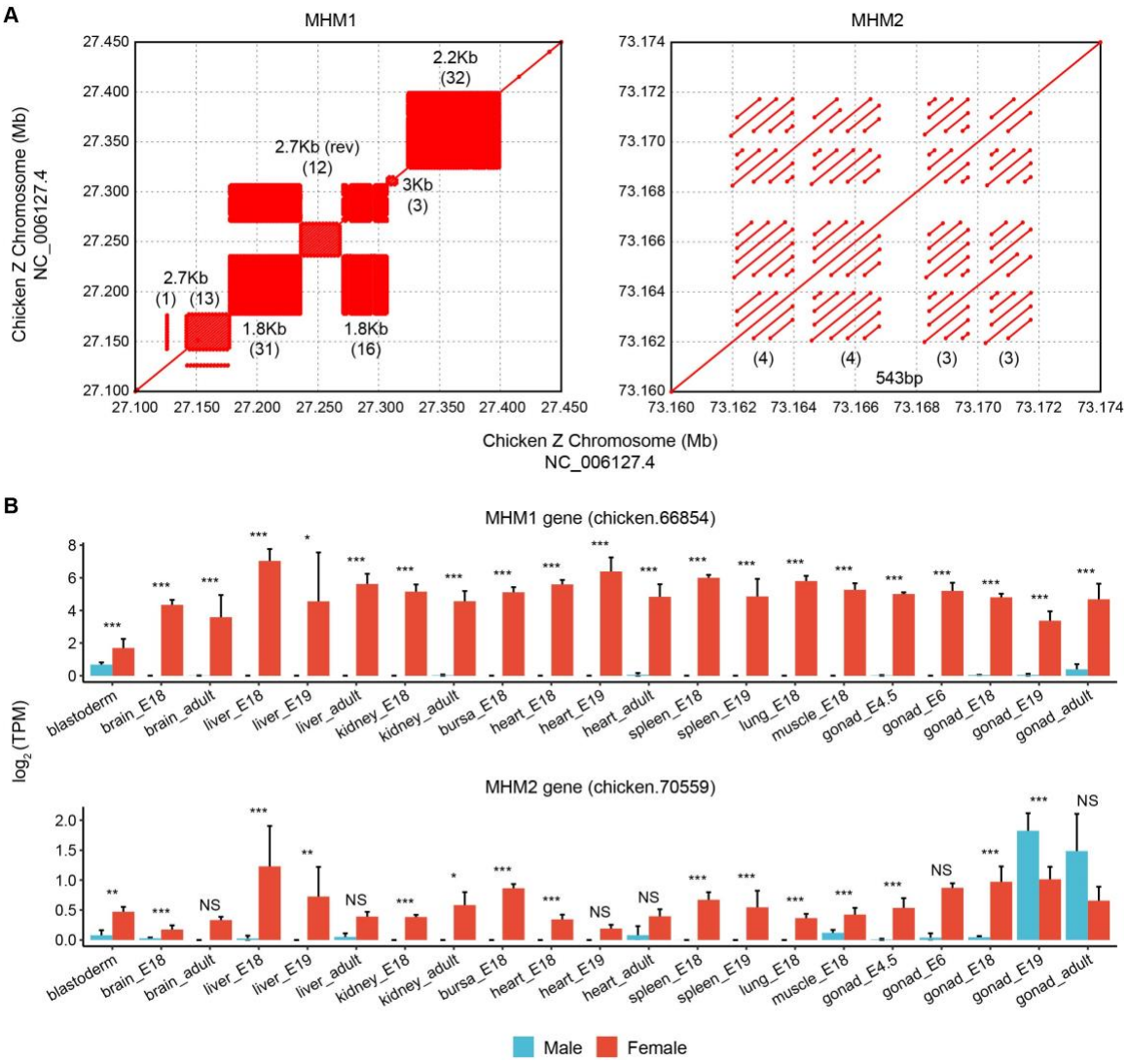

Figure 2.6 Repetitive nature and lncRNA expression patterns of MHM1 and MHM2 loci. **(A)** Both loci are highly repetitive. Each rectangular area filled with matches in dotplots depicts a block of tandem repeats. The length of each repeat unit and the number of iteration time (in parentheses) are shown. **(B)** Expression differences between males and females for lncRNAs transcribed from MHMs. 'E' is short for 'embryonic day'. For each locus, one example lncRNA gene with the highest average expression is shown. Significant expression differences between males and females were tested using DESeq2 with raw counts generated from StringTie (***: $Q < 0.001$; **: $Q < 0.01$; *: $Q < 0.05$; NS: not significant).

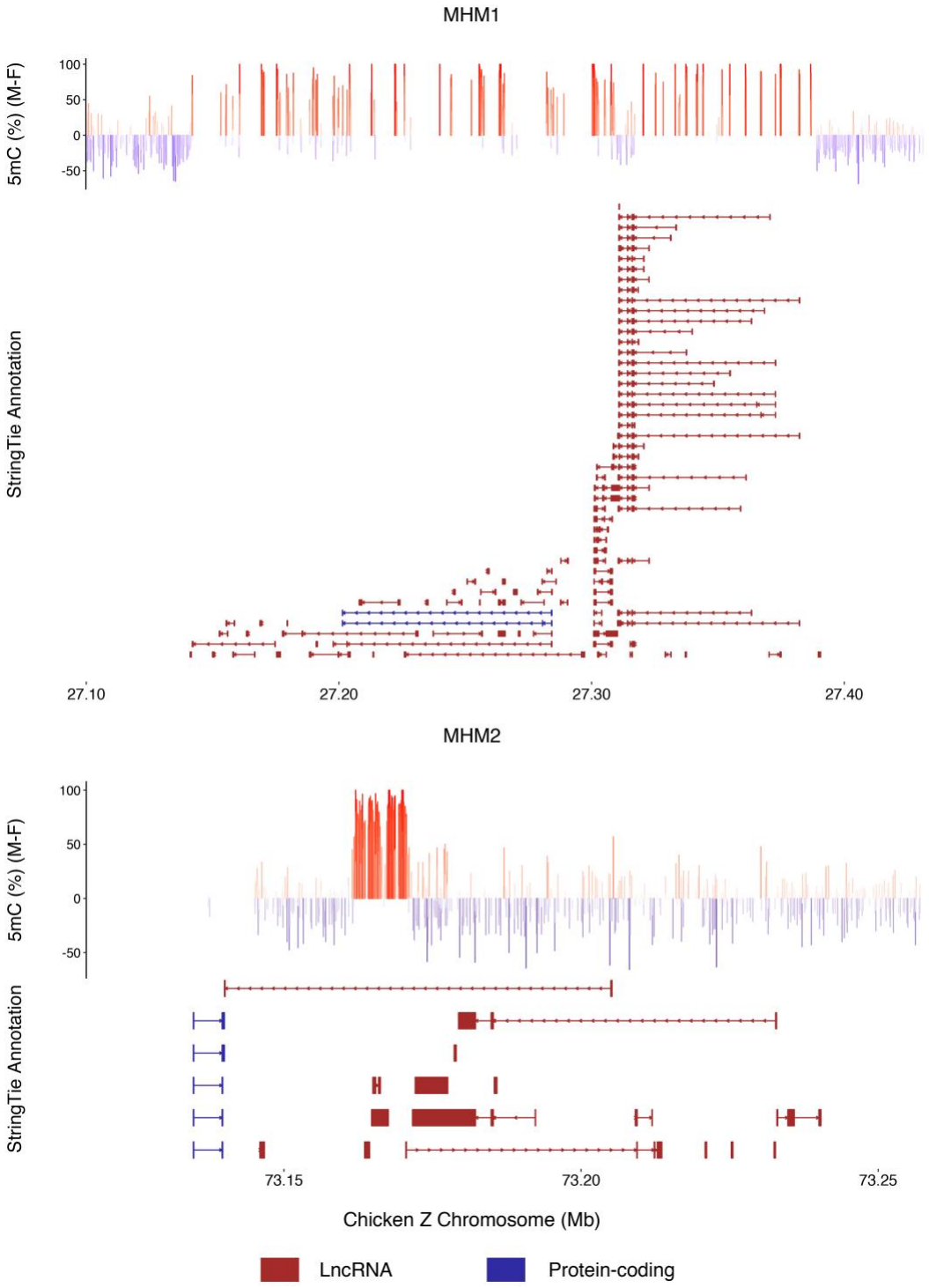

Figure 2.7 StringTie annotation of genes within or near the two MHM loci. The vast majority of genes within or near the two MHM loci are lncRNA genes. Coding potential was predicted using FEELnc (Wucher et al. 2017), and transcripts with coding potential greater than 0.4214 (a cutoff chosen by the program) were defined as protein-coding genes. To show the borders of MHMs, we plotted differences in DNA methylation between male and female lung samples on the upper panel for each locus. Each line represents a CpG site.

Despite these similarities, MHM2 seemed to have a broader phylogenetic distribution than MHM1. Specifically, using an E-value cutoff of 10^{-5} as the BLAST criteria (Methods), we found the repeat unit of MHM2 in Galloanserae (both waterfowls and landfowls) and in multiple orders of Neoaves (Table 2.1). However, MHM2 was only found repetitive within the Galloanserae lineage (Table 2.1). The absence of MHM2 sequence in Palaeognathae or other Neoaves lineages (e.g., passerines) suggested that MHM2 might have originated in the ancestor of Neognathae birds but was subsequently lost in some Neoaves. Due to the lack of DNA methylation data of Neoaves species harboring the MHM2 sequence, we could not test whether the sex-specific methylation status had a more ancient origin.

Table 2.1 Species that potentially harbor the repeat-unit sequence of MHM2.
The MHM2 repeat unit was blasted against the database of RefSeq representative genomes of birds (taxid:8782) using BLASTn (under the "blastn" mode). This table lists all species with hits satisfying E-value $<10^{-5}$.

Superorder	Order	Species (Scientific Name)	Species (Common Name)	BLAST E-value	Repetitive? (# of iterations)
Galloanserae	Galliformes	*Gallus gallus*	Red junglefowl	0	Yes (14)
		Coturnix japonica	Japanese quail	4×10^{-141}	Yes (27)
		Meleagris gallopavo	Wild turkey	5×10^{-139}	Yes (>=3; Poor assembly quality)
		Numida meleagris	Helmeted guineafowl	6×10^{-18}	Unknown (>=1; Poor assembly quality)
	Anseriformes	*Anser cygnoides domesticus*	Domestic goose	9×10^{-10}	Yes (>=2; Poor assembly quality)
		Anas platyrhynchos	Mallard	2×10^{-06}	Yes (28)
Neoaves	Accipitriformes	*Aquila chrysaetos canadensis*	Golden eagle	8×10^{-17}	No (1)
		Haliaeetus leucocephalus	Bald eagle	8×10^{-17}	No (1)
		Haliaeetus albicilla	White-tailed eagle	8×10^{-17}	No (1)
	Falconiformes	*Falco cherrug*	Saker falcon	3×10^{-09}	No (1)
		Falco peregrinus	Peregrine falcon	3×10^{-09}	No (1)
	Procellariiformes	*Fulmarus glacialis*	Northern fulmar	3×10^{-15}	No (1)
	Pelecaniformes	*Pelecanus crispus*	Dalmatian pelican	3×10^{-10}	No (1)
		Phaethon lepturus	White-tailed tropicbird	9×10^{-10}	No (1)
	Gruiformes	*Chlamydotis macqueenii*	Houbara bustard	3×10^{-10}	No (1)
	Charadriiformes	*Charadrius vociferus*	Killdeer	1×10^{-08}	No (1)

2.2.4 Reduced male-to-female expression ratios of genes neighboring MHM1 and MHM2

Although the global the male-to-female expression ratios across the avian Z chromosomes is near 1.5 (Ellegren et al. 2007; Itoh et al. 2007; Graves 2016), several previous studies showed that genes located near MHM1 displayed pronounced reduction in male-to-female expression ratios in chicken and other Galloanserae species (Melamed and Arnold 2007; Mank and Ellegren 2009). Using the above large-scale RNA-seq data, we found that neighboring genes of both MHM1 and MHM2 exhibited reduced male-to-female ratios compared to the rest of the Z chromosome in most somatic and gonadal tissues across developmental stages (Figure 2.8). By performing a changepoint analysis to detect abrupt changes in ratios across the Z chromosome (Methods), we found that the boundary for MHM1-affected genes was 25 – 32 Mb and the boundary for MHM2-affected genes was 72.5 – 73.5 Mb.

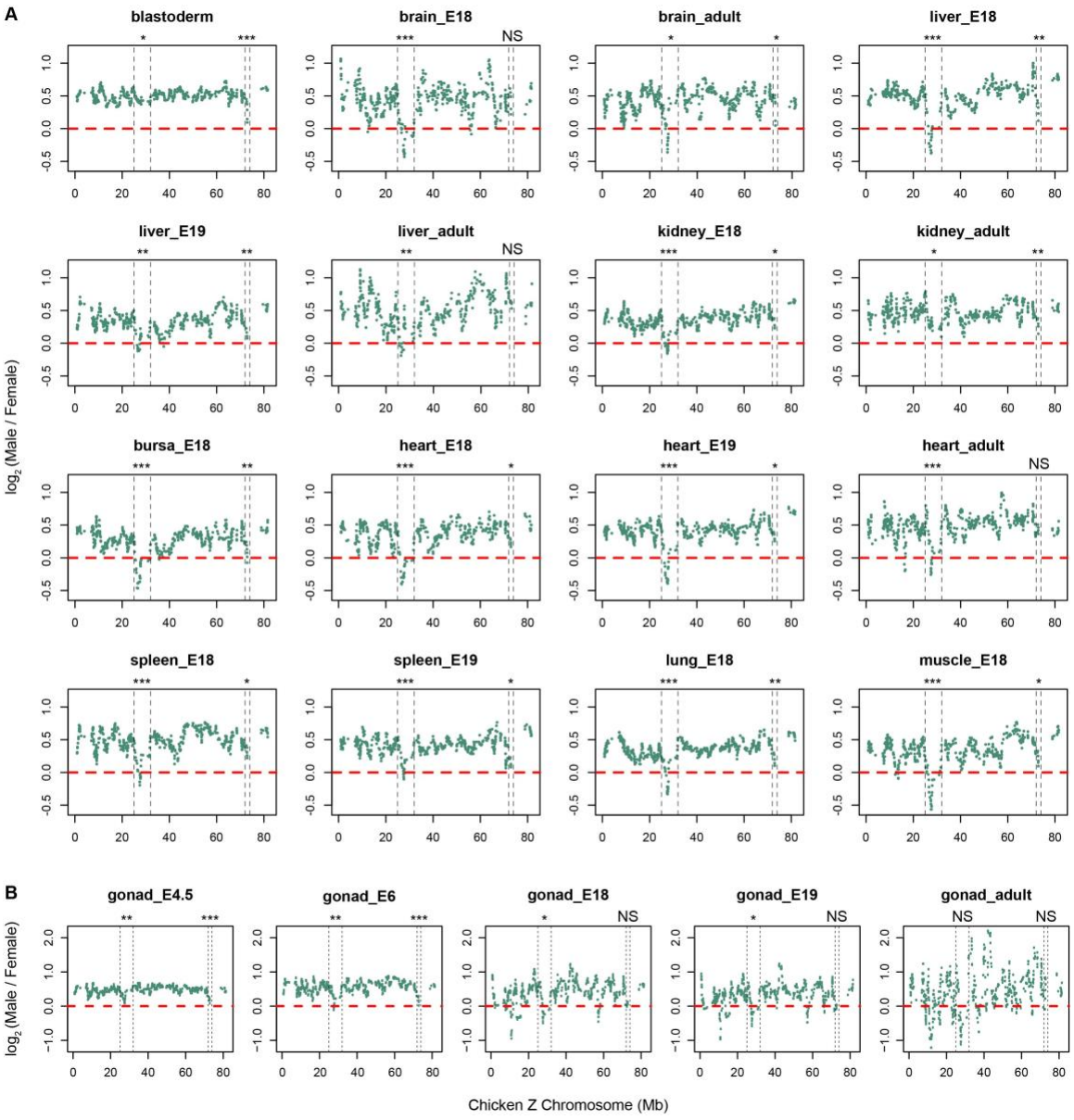

Figure 2.8 The $\log_2$ (Male/Female) values across the Z Chromosome. Results are shown with 10-gene windows for **(A)** somatic and **(B)** gonadal tissues. 'E' is short for 'embryonic day.' The dashed red lines (zeros) represent the no dose bias towards either sex. Genes with average log2(TPM) lower than 1 in either males or females were filtered out. The dashed black lines depict the boundaries of potential MHM1- or MHM2-affected protein-coding genes identified by a changepoint analysis (Methods). For MHM1, the boundary is 25 – 32 Mb, and for MHM2, 72.5 – 73.5 Mb. One-tailed Mann–Whitney U tests were used to test whether the MHM1 or MHM2 neighboring region consists of genes with lower Male/Female ratios than the Z background (***: $P < 0.001$; **: $P < 0.01$; *: $P < 0.05$; NS: not significant).

2.2.5 How did the sex difference in DNA methylation and expression of genes at and near MHMs originate?

The next question we asked was whether the localized differences in DNA methylation between male and female Z chromosomes was caused by male hypermethylation (reduced male expression of neighboring genes) or female hypomethylation (increased female expression of neighboring genes) or both. This question is difficult to answer for the unknown ancestral methylation and expression status at or near MHMs. However, we borrowed the idea from previous studies in which the authors inferred the expression of proto-sex chromosomes using autosomal orthologs from outgroup species (Brawand et al. 2011; Julien et al. 2012; Mank 2013; Cortez et al. 2014; Marin et al. 2017). For instance, the expression of chicken autosomal orthologs was used to represent the ancestral expression of human and lizard X-linked genes (Julien et al. 2012; Marin et al. 2017).

In the current study, we used available brain RNA-seq data from two divergent outgroup species—blue tit (a passerine in Neoaves) and ostrich (in Palaeognathae) (Adolfsson and Ellegren 2013; Mueller et al. 2015), both lacking MHM loci, to infer the ancestral expression levels of MHM neighboring genes. We matched the evolutionary history for each MHM locus in cross-species comparisons. That is, for MHM1 which is located in the oldest stratum (S0), both outgroup lineages were used (Wang et al. 2014; Zhou et al. 2014); in contrast, as MHM2 was located in a younger stratum (S1) shared by only chicken and blue tit (in ostrich this region is located in the pseudoautosomal region) (Wang et al. 2014; Zhou et al. 2014), only blue tit was used as an outgroup.

By comparing the MHM1-neighboring genes among all three species, we found strong evidence for up-regulation in female chickens (Figure 2.9A-C), implying

hypomethylation of MHM1 in females (hypomethylation -> increased transcription of lncRNAs -> overexpression of neighboring genes in females). In contrast, the comparison between chicken and blue tit for MHM2 yielded inconclusive results (Figure 2.10).

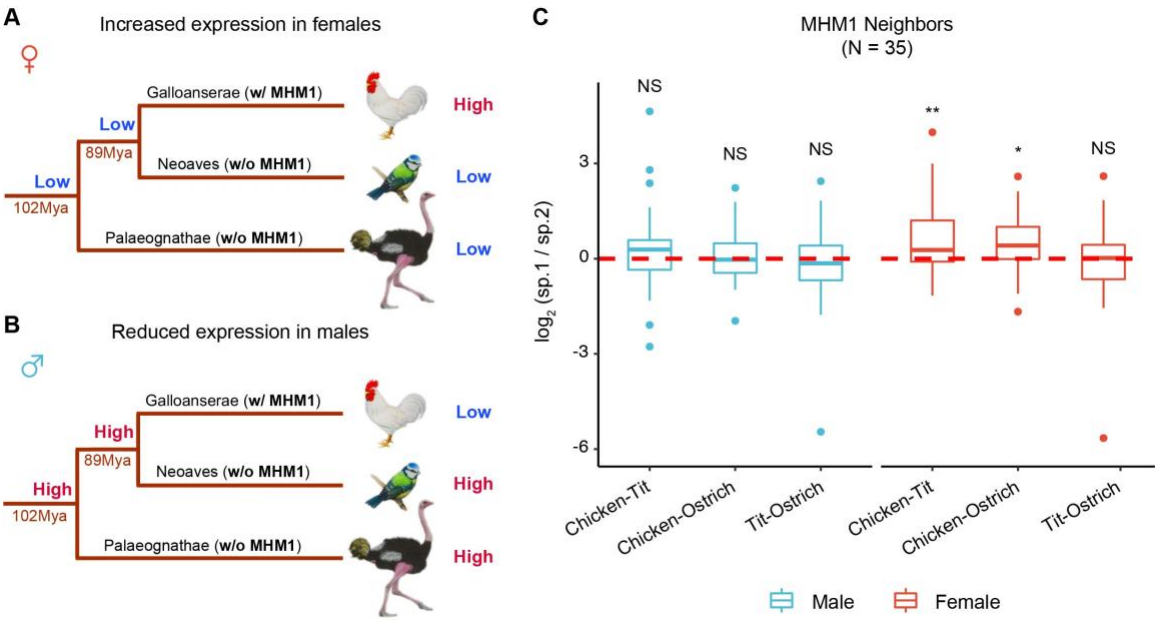

Figure 2.9 Are reduced male-to-female expression ratios of MHM1 neighboring genes due to female upregulation or male downregulation? (A-B) Alternative scenarios to achieve reduced male-to-female expression ratios. A pattern similar to (A) depicts up-regulation in female chickens, while (B) will support down-regulation in male chickens. 'High' and 'Low' describe the levels of gene expression in a species or in a common ancestor. The marked expression levels in common ancestors are based on the parsimony method. **(C)** Pairwise species differences in expression of the MHM1 neighboring genes (adult brain). For each gene, expression was averaged across samples per species, and the pairwise expression difference (the ratio of expression of species 1 to the expression of species 2) was $\log_2$-scaled. Only orthologs present in all species were used (numbers of genes compared are in parentheses). The statistical significance was evaluated using paired Mann–Whitney U tests (**: $P < 0.01$; *: $P < 0.05$; NS: not significant).

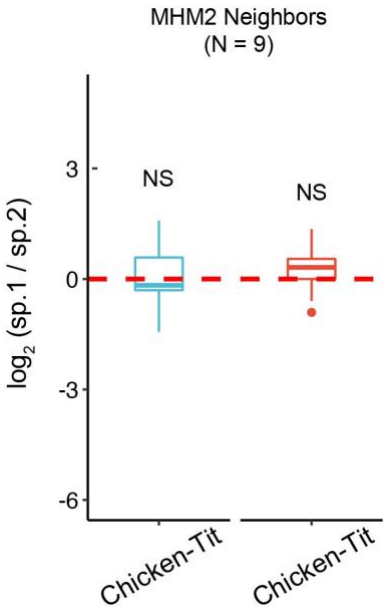

Figure 2.10 Pairwise species differences in the expression of protein-coding genes near MHM2. For each gene, expression was averaged across samples per species, and the pairwise expression difference (the ratio of expression of species 1 to the expression of species 2) was $\log_2$-scaled. Only orthologs present in both species were used (numbers of genes compared are in parentheses). The statistical significance was assessed using paired Mann–Whitney *U* test (NS: not significant).

2.3 Discussion

In eutherian mammals, the non-recombining Y chromosome triggers dosage compensation that is achieved by inactivation one of the X chromosomes in females (Carrel and Willard 2005; Graves 2016). Previous studies have established that epigenetic mechanisms are critical to this process (Hellman and Chess 2007; Brockdorff and Turner 2015; Lucchesi and Kuroda 2015). Birds, however, exhibit incomplete dosage compensation of the Z chromosome, irrespective of the extent of W degeneration that vary

across avian lineages (Ellegren et al. 2007; Itoh et al. 2007; Adolfsson and Ellegren 2013). Therefore, birds offer a unique opportunity to study how epigenetic modifications and dosage compensation are associated in general.

Here, we used single-nucleotide-resolution DNA methylation maps in the human, chicken (Galloanserae) and white-throated sparrow (Neoaves) to examine how the X/Z chromosome is epigenetically differentiated between males and females in these species. Although the human X chromosome shows moderate sex differences in DNA methylation, the Z chromosome in both avian species have similar DNA methylation levels between sexes, which is consistent with the lack of global dosage compensation in birds. Despite the global similarity, two regions are highly differentiated in their DNA methylation patterns between males and females in the chicken, but not the white-throated sparrow (Figure 2.3; Figure 2.4). The first region corresponds to a previously described locus identified via a targeted approach (referred to as 'MHM1' in our study) (Teranishi et al. 2001), and it has long been considered to be the only region with high differentiation in DNA methylation between sexes. Here, we identified a novel region on the Z chromosome (referred to as 'MHM2'). We found little sequence similarity between the two MHM loci, indicating their independent evolutionary origins. MHM1 sequences are present in Galloanserae and absent in the other two major avian lineages, while the MHM2 sequence is present in Galloanserae and several Neoaves lineages, which is consistent with its more ancient origin (Table 2.1). Nevertheless, several striking features are shared between MHM1 and MHM2: 1) females exhibit higher chromatin accessibility in MHMs (Figure 2.3; Figure 2.5). This finding is consistent with previous studies of MHM1 that detected the enrichment of H4K16ac near this locus (Bisoni et al. 2005; Itoh et al. 2010; Itoh et al. 2011). 2) both loci are transcribed into lncRNAs in a female-biased manner (Figure 2.6B; Figure 2.7); 3) although the Z chromosome is generally male-biased, MHM-neighboring genes display

reduced male-to-female expression ratios (Figure 2.8); 4) both loci are composed of tandem repeats (Figure 2.6A); These convergent features between the two MHMs potentially indicate that epigenetic regulation of the two loci evolved by similar molecular mechanisms.

Although the direct link between MHMs and expression of neighboring genes requires further experimental validation, the reduced male-to-female ratios of genes near both MHM1 and MHM2 are consistent with the influences of MHMs on the transcription of these genes. This model is further supported by the observation that the region exhibiting reduced male-to-female ratios near MHM2 (~1 Mb) appears narrower than the one near MHM1 (~7 Mb), which could result from the fewer number of transcripts and lower expression of MHM2 lncRNAs relative to MHM1 lncRNAs (Figure 2.6B; Figure 2.7). Therefore, the previous model in which long non-coding RNA (lncRNA) from MHM1 bind to neighboring genes and increase their expression in female chicken (Melamed and Arnold 2007; Mank and Ellegren 2009; Wright et al. 2015) can also be applied to MHM2 (Figure 2.11).

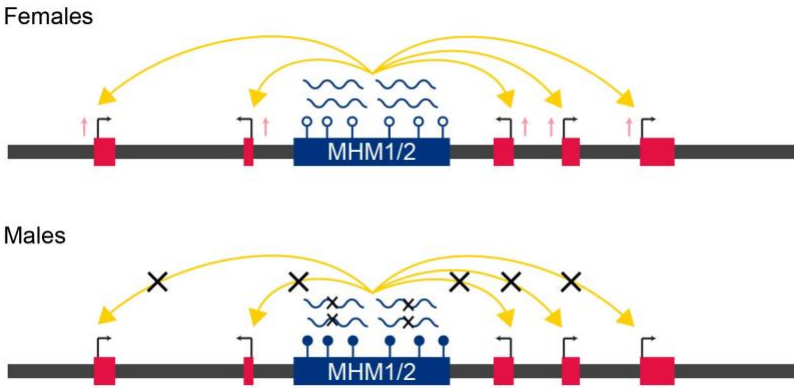

Figure 2.11 The current model of the effects of MHMs on neighboring genes. In females, CpGs are hypomethylated (empty circles) in MHM1 and MHM2, and lncRNAs are transcribed. These lncRNAs bind neighboring genes to up-regulate their expression. In contrast, CpGs are hypermethylated (solid circles) in MHMs in males, and the transcription of lncRNAs is suppressed. Accordingly, the MHM loci have no effects on neighboring genes.

The functional significance of tandem repeat blocks in the placental *XIST* and marsupial *Rsx* lncRNAs (independently evolved master regulators of X inactivation) (Grant et al. 2012; Brockdorff 2018) may provide more clues about why both MHMs have large arrays of repetitive sequences. For example, the tandem repeats in *XIST* have been shown to bind to RNA binding proteins (RBPs) to mediate *XIST* localization and silencing (Chu et al. 2015; Ridings-Figueroa et al. 2017). The expansion of tandem repeats may create more RBP binding sites and increase the efficiency of dosage compensation (Brockdorff 2018). For MHMs, the long-range impacts on neighboring genes (MHM1: 7 Mb; MHM2: 1 Mb) are likely to be facilitated by the amplification of the functional units.

We sought to test whether the reduction of male-to-female expression ratios of MHM nearby genes is due to a reduction of expression in males or to an increase of expression in females. Here, we adopted a similar approach using outgroup species (Brawand et al. 2011; Julien et al. 2012; Mank 2013; Cortez et al. 2014; Marin et al. 2017)

to compare the expression of current Z-linked genes (with MHMs) with the expression of proto-sex chromosomes (without MHMs). The results were consistent with female up-regulation of MHM1-neighboring genes (and hence MHM1 hypomethylation in females) but inconclusive for MHM2-neighboring genes (Figure 2.9; Figure 2.10). Because the chicken genome, similar to other vertebrate genomes, is globally heavily methylated (Elango and Yi 2008; Suzuki and Bird 2008), hypomethylation of MHM1 may be considered a 'gain of function'.

When is the sex differences in chromatin configuration of the two MHM loci established? Teranishi et al. (2001) examined early chicken embryos and found that the hypermethylation of MHM1 in males relative to females was already established in E1 embryos. In this study, we examined blastoderms at 12-hr, which represents the earliest developmental stage at which chicken embryos are accessible (Ayers et al. 2013). Although other tissues have much lower expression of MHM1 in males than in females (log_2-scaled male-to-female ratios > 4), the higher expression of MHM1 in males and lower expression in females relative to later developmental stages lead to milder differences between sexes in the blastoderm stage (log_2-scaled male-to-female ratios < 1) (Figure 2.6B). Accordingly, the effect of MHM1 is the weakest at this stage (Figure 2.8). These observations may suggest that 1) male and female differences in methylation of MHM1 are established right after fertilization; 2) both hypermethylation in males and hypomethylation in females of MHM1 are gradual processes and complete later in development. In contrast, MHM2 seems to exert its effects very early, evidenced by a strong dosage valley in the blastoderm (Figure 2.8). However, the higher expression of MHM2 in males and the absence of MHM2 valleys in gonads at later developmental stages suggest that MHM2's function in gonads is time-dependent (Figure 2.6; Figure 2.8). For both MHM loci, it would be of interests to use epigenetic and transcriptomic data from

multiple time points to understand the exact timing of the sex differences in DNA methylation and when the differences are stabilized (or disappear).

What drives the evolution of regional epigenetic differentiation of the male and female Z chromosomes? One hypothesis was that MHM1 was involved in sexual differentiation by silencing the candidate avian sex-determining gene *DMRT1*, which is in close proximity to MHM1, in females (Teranishi et al. 2001). Nevertheless, we found that similar to other genes within the MHM1 valley, *DMRT1* is expressed at similar levels between sexes in the majority of somatic tissues, and even significantly female-biased in samples from brain (E18) (Figure 2.12). In gonadal tissues across embryonic stages, although *DMRT1* is significantly male-biased, the male-to-female expression ratio is near 2 (log_2-scaled male and female differences ≈ 1) (Figure 2.12), potentially arising from the difference in the number of Z chromosomes between the sexes (male: ZZ; female: Z). These results seem to conflict with the hypothesis and indicate limited silencing effects of MHM1 on *DMRT1* in females for gonadal tissues at the embryonic stage and most somatic tissues. However, as suggested by an earlier study, MHM1 may exert its effects on *DMRT1* only at the post-transcriptional level (Roeszler et al. 2012). Despite little effect of MHM1 on *DMRT1* transcription in the gonad of embryos, *DMRT1* expression is strongly male-biased in adult gonads (log_2-scaled male and female differences ≈ 6) (Figure 2.12), in agreement with potential down-regulation in females. These observations suggest that the role of *DMRT*1 in sex determination in chicken is possibly developmental-stage- and tissue-dependent and needs to be further evaluated.

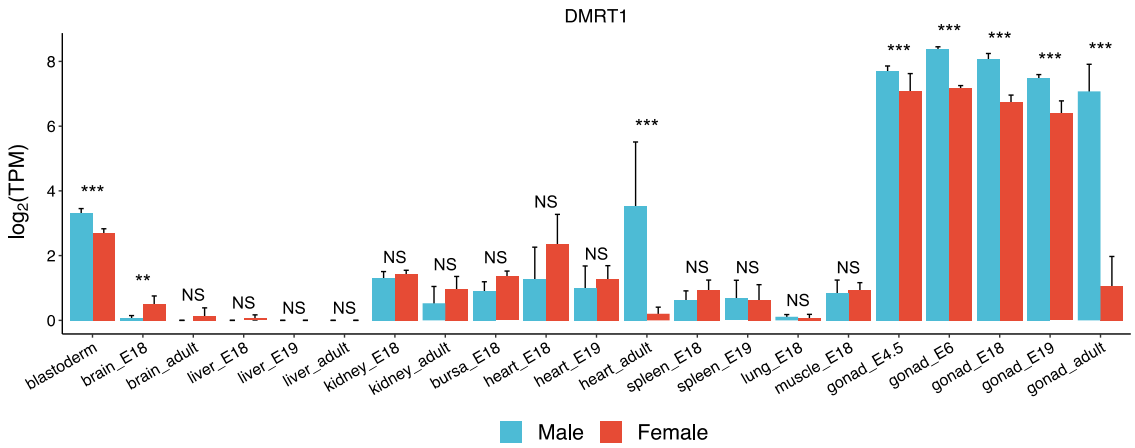

Figure 2.12 Expression differences in *DMRT1* between males and females. Significant expression differences between males and females were tested using DESeq2 with raw counts generated from StringTie (***: $Q < 0.001$; **: $Q < 0.01$; NS: not significant).

Another hypothesis for the evolution of MHM loci is their critical regulation of neighboring genes with functions critical to females (Mank and Ellegren 2009; Wright et al. 2015). Even though we and others (e.g., Wright et al. 2015 for MHM1) have not identified any functional enrichment, some of the genes located nearby MHMs include some known to have highly female-specific functions. For example, *VLDLR* (encodes very-low density lipoprotein receptor), which is near MHM1, is involved in oocyte growth via yolk deposition in hens (Bujo et al. 1997). Additionally, *RLN3* (encodes relaxin 3) near MHM1 and *CARTPT* (encodes cocaine- and amphetamine-regulated transcript) near MHM2 are both female-biased in the brain data and are shown to have female-biased effects in response to food intake in rodents (Asnicar et al. 2001; Lenglos et al. 2015). However, such effects have not been reported for birds. However, the exact function of MHM loci needs experimental validation.

In this study, we contrast epigenetic differentiation of the X and Z chromosomes between sexes and link the findings to different extents of dosage compensation in male- and female-heterogametic systems. Extremely localized sex differences in DNA methylation on the Z chromosome evolved at least twice in the Galloanserae lineage, suggesting their potentially essential function in sex differentiation and dosage compensation.

2.4 Methods

2.4.1 Enhanced chicken genome assembly

We used an enhanced Gallus_gallus-5.0 (GCA_000002315.3) assembly for our analyses. Specifically, the W chromosome was replaced by the one from the latest GRCg6a (GCA_000002315.5) assembly (Bellott et al. 2017) for its improved quality (Gallus_gallus-5.0: 5.16 Mb vs. GRCg6a: 6.81 Mb).

2.4.2 Whole-genome bisulfite sequencing data and DNA methylation calling

To compare DNA methylation between Z and W gametologs and to compare DNA methylation between males and females, we collected previously published chicken WGBS data sets across four tissues (brain, retina, lung, and muscle) (Li et al. 2015; Lee et al. 2017; Zhang et al. 2017) (Table A. 1), which comprises either samples from both sexes (brain, retina, and lung) or from females alone (muscle). Additionally, to compare the sex differences in methylation between chicken and other species from Neoaves, we performed brain WGBS of the white-throated sparrow (tan morph) for a male adult and a female adult. We also collected brain WGBS data from a male great tit adult to determine the relationship between gene expression and DNA methylation (Laine et al. 2016).

WGBS reads were trimmed with TrimGalore 0.4.5 using a quality cutoff of 30 and then aligned to the enhanced assembly for chicken samples (or to the 'Parus_major1.1' assembly for the great tit sample) using Bismark v0.18.1 in bowtie2 mode (Krueger and Andrews 2011). Subsequently, mapped reads were deduplicated using deduplicate_bismark, non-bisulfite-converted reads were filtered using filter_non_conversion (with the --percentage_cutoff 20 option), and fractional methylation was extracted using bismark_methylation_extractor. Methylation data were then merged across samples for each sex. Only CpGs with at least three mapped reads were retained for further analysis.

2.4.3 DNA methylation comparison between Z and W gametologs

In total, 26 Z-W pairs were annotated in NCBI (W sequence is from the latest GRCg6a version) (**Table A. 2**). Since the transcription start site (TSS) of W gametologs may be poorly annotated, promoters of Z gametologs (defined as -1.5 kb – 500bp of TSS) were blasted against the W sequence to identify homologous promoters on W by the blastn program in BLAST 2.7.1+ (-task blastn). Gene bodies or promoters of Z-W pairs were then aligned using MAFFT v7.407. Only aligned CpGs with more than three reads covered for both Z and W gametologs were retained for further analysis.

2.4.4 Identification of outliers for sex differences in methylation

To identify outlier regions of sex differences in DNA methylation across the whole chicken genome, we computed DNA methylation levels averaged across CpGs using a 10 Kb window size and a 1 Kb step size (a threshold required that at least 20 CpGs with methylation data be present in a window). The differences between males and females were then calculated. We identified extreme outliers which were defined as windows with

differences in methylation between sexes outside the range of the first quartile minus three times the interquartile range and the third quartile plus three times the interquartile range (i.e., $< Q_1 - 3 * IQR$ or $> Q_3 + 3 * IQR$). To identify potentially non-tissue-specific outlier regions, we retained outlier windows common to the brain, retina, and lung. Except for MHM1 and MHM2 on the chicken Z Chromosome, we did not identify outliers for sex differences in methylation on any autosomes (macrochromosomes are shown as examples in **Figure A. 1**).

2.4.5 ATAC-seq data, signal normalization, and differential peak identification

We collected recently published ATAC-seq data from the liver, $CD4^+$ T cells (Foissac et al. 2018) and flight muscles/bones (Sackton et al. 2018) (Table A. 1). ATAC-seq reads were trimmed with TrimGalore 0.4.5 and a quality cutoff of 30, and the processed reads were aligned to the enhanced chicken genome using bowtie2 2.3.4.2 (Langmead and Salzberg 2012) with -X 2000 --no-mixed --no-discordant parameters. Aligned reads were then merged across samples for each sex. Normalized ATAC-seq signals were obtained by running callpeak from the MACS2 2.1.1.20160309 program (Zhang et al. 2008) with the -p 0.01 -B --nomodel --SPMR --shift -100 --extsize 200 options. The signals (fragments pileup per million reads) for each sex at the two MHM loci were plotted with ggbio 1.28.5. To find statistically significant ATAC-seq peaks, we reran callpeak without --SPMR and identified differential peaks using bdgdiff.

2.4.6 RNA-seq data, transcriptome assembly, and lncRNA identification

RNA-seq data from 165 chicken samples across multiple somatic tissues (blastoma, brain, bursa of Fabricius, heart, kidney, liver, lung, muscle, spleen) and developmental stages (12 hrs, E18, E19 and adults) were collected from multiple sources

(Brawand et al. 2011; Julien et al. 2012; Muyle et al. 2012; Ayers et al. 2013; Uebbing et al. 2015; Zimmer et al. 2016; Marin et al. 2017) (**Table A. 1**). Raw reads were trimmed with TrimGalore 0.4.5 and then aligned to the enhanced chicken genome using HISAT2 (Kim et al. 2015). We filtered out secondary alignments using SAMtools 1.7 (Li et al. 2009) to ensure that only primary alignments were retained for further estimation of gene expression.

To identify lncRNA genes within MHM loci and quantify the transcripts, we used StringTie 1.3.4d (Pertea et al. 2015) to assemble aligned reads into transcripts. Then, the coding potential of these transcripts was predicted using FEELnc v.0.1.1 (Wucher et al. 2017) under the *shuffle* mode, and transcripts with the coding potential score smaller than 0.4214 (a cutoff chosen by the program) were defined as lncRNAs. It should be noted that since our collected RNA-seq datasets are poly(A)-selected, we could capture only polyadenylated lncRNAs.

2.4.7 Quantification of gene expression and differential expression analysis

We measured the expression of these lncRNAs by rerunning StringTie against the annotation of the new transcript assembly with the -e -b options. Meanwhile, we quantified protein-coding genes on the Z chromosome by running StringTie against the chicken Ensembl annotation (Gallus_gallus-5.0.92 release) with the same options. The Log_2-scaled Transcripts Per Kilobase Million (TPM) measure (specifically, $\log_2(TPM + 1)$) was chosen to represent gene expression levels for further analysis.

We averaged expression across samples within each sex, and genes with average expression above 1 in either sex were retained. The running average (window size of 10 genes with a slide size of one gene) of the male-to-female ratios on the Z chromosome

was computed with gtools 3.8.1. Expression valleys (dips in male/female ratios) near MHM1 and MHM2 were tested using a one-tailed Mann–Whitney *U* test.

Additionally, we performed differential expression between males and females using DESeq2 1.18.1 (Love et al. 2014), using raw counts generated with prepDE.py from the StringTie package.

2.4.8 Identification of repeat units in MHM loci

We detected tandem repeat units of MHM1 and MHM2 using XSTREAM (Newman and Cooper 2007). We identified five different repeat units of MHM1 -- 2.7Kb, 1.8Kb, 2.7Kb_rev (reverse complement of the 2.7Kb unit), 3Kb and 2.2Kb (Figure 2.6). Among them, the 2.7Kb_rev (JF692775) and 1.8Kb units (JF692776, named 'sMHM-BamHI') were previously identified in (Itoh et al. 2011), and the 2.2Kb unit was discovered by Teranishi et al. (2001) (AB046698). As described by Itoh et al. (2011), the five repeat units share a core sequence (JF692776: 957-1219) which is potentially important for their function. For MHM2, a 543bp unit is iterated 3-4 times per repeat block (Figure 2.6). All repeat units of MHM loci are relatively GC rich (GC content: 55-60%), while the GC content for the chicken Z chromosome (the Gallus_gallus-5.0 assembly) is 40.74%.

2.4.9 Detection of the boundaries of MHM-affected genes

To obtain accurate boundaries of MHM-affected genes (dosage valleys), we performed a changepoint analysis using the R package 'changepoint.np' (Haynes et al. 2016). This package can identify multiple change-points in time series-like data, without any assumptions regarding the distribution. The changepoints across tissues/developmental stages were merged to obtain all potentially affected genes near MHM1 or MHM2.

2.4.10 Cross-species expression comparison

We obtained RNA-seq data from adult brains of the great tit (Laine et al. 2016), the blue tit (Mueller et al. 2015) and the ostrich (Adolfsson and Ellegren 2013) (Table A. 1). The reads were pre-processed, and expression levels (in $\log_2$(TPM + 1)) were measured using the same pipeline as in chicken, except that reads were aligned to the 'Parus_major1.1' assembly for the great tit sample, to the 'cyaCae2' assembly for blue tit samples, and to the 'ASM69896v1' assembly for ostrich samples.

To identify cross-species orthologs (chicken - blue tit - ostrich), we ran Proteinortho V5.16b (Lechner et al. 2011) on the protein sequences of three species (NCBI annotations for the blue tit and ostrich, and Ensembl annotation for the chicken), with -synteny -identity=50 -cov=50 -selfblast options. Using this method, we identified 9,991 orthologs present in chicken, blue tit and ostrich.

To ensure comparability of expression across samples of different species, we used a median scaling method (Brawand et al. 2011; Julien et al. 2012; Marin et al. 2017). In detail, for genes with expression that are consistently within the inner quartile range (2,111 genes across all 23 samples), the median expression per sample was computed. We then scaled the medians to a common value (mean of all medians) by scaling factors, and the scaling factor was used to calculate the new expression value for all genes in each sample.

2.4.11 Sample sexing

Since the sexes of several WGBS (brain and retina) and ATAC-seq (flight muscles/bones) samples were unknown, we sexed the samples based on the percentage of reads that could be mapped to the W chromosome out of all aligned reads (W%). The

distribution of percent W reads revealed two clear groups corresponding the two sexes (**Figure A. 2**), and we determined the sex based on this stratification.

2.4.12 The presence or absence of MHM repeat units in other avian species

We blasted all MHM repeat units against all representative genomes of birds (taxid: 8782) in NCBI (last access: 4/28/2019) using the Nucleotide BLAST in "blastn" mode. This mode is optimized for cross-species search. We set the threshold of E-value to 10^{-5}. Although both MHM1 and MHM2 can be identified in chicken and other Galloanserae species, MHM2 is also present in multiple orders of Neoaves (Table 2.1).

To validate the presence or absence of MHM2 in other avian species, we aligned the Z chromosome of these species to the chicken Z chromosome using LASTZ 1.04.00 (Harris 2007) with the --notransition --nogapped --step=20 --rdotplot options. The output dotplots were zoomed in to MHM2 neighboring regions and visualized in R (R Core Team 2019).

2.4.13 Self-alignment of MHM loci

We self-aligned each MHM locus to test for repetitive sequences. Specifically, nucmer from the MUMmer 4.0.0beta2 (Marcais et al. 2018) was run for an MHM locus against itself with --maxmatch --nosimplify options. We then used mummerplot to draw dot plots for these self-alignments. The two MHM loci were also aligned against each other using the same approach, but no alignment was generated.

2.4.14 Gene ontology analysis

PANTHER Overrepresentation Test (Released 20171205) from the PANTHER Classification System version 11 (Mi et al. 2017) was used to test the enrichment of gene

ontology for MHM1 or MHM2 neighboring genes. Ensembl gene IDs were used for these genes, and the reference list used was the *Gallus gallus* annotation (Panther annotation version 13.1). We found that no gene ontology terms from the biological process or molecular function were enriched in these sets of genes.

2.5 Acknowledgements

This work was supported by grants from National Science Foundation (IOS 1656247) and support from the Elizabeth Smithgall Watts endowment and the Georgia Tech School of Biological Sciences to Dr. Soojin Yi, and an NIH grant (R01MH082833) to Dr. Donna Maney and Dr. Soojin Yi. We thank the Yi laboratory for comments on the manuscript and Xiulan Pan for help with the illustration.

CHAPTER 3

REGULATORY DIVERGENCE AND DOSAGE COMPENSATION OF SEX CHROMOSOME-LIKE AUTOSOMES IN THE WHITE-THROATED SPARROW

3.1 Introduction

A pair of autosomes in the white-throated sparrow (*Zonotrichia albicollis*) shares some similarities with sex chromosomes. This North American songbird features a striking phenotypic dimorphism: birds of the white-striped (or "white") morph display more territorial aggression and invest less parental care than their tan-striped (or "tan") conspecifics of the same sex (Figure 3.1). The phenotypic differences can be traced to a large chromosomal rearrangement harboring ~1,000 genes on the second chromosome, which occurred approximately 2-3 million years ago (Thorneycroft 1975; Thomas et al. 2008; Huynh et al. 2010). Tan birds are homozygous for the standard version (ZAL2/ZAL2), and white birds are heterozygous (ZAL2/ZAL2^{m}) (Thorneycroft 1975) (Figure 3.1). Interestingly, birds of the distinct morphs mate exclusively with each other at a >98% frequency (Tuttle et al. 2016). The strong disassortative mating preference leads to suppressed recombination of the rarely homozygous ZAL2^{m} chromosome.

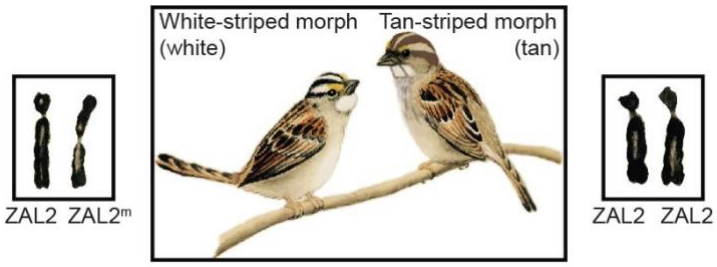

Figure 3.1 Two plumage morphs of the white-throated sparrow, with their second chromosome karyotypes shown in the boxes.

Previously, cytogenetic studies by Thomas et al. (2008) and Davis et al. (2011) revealed that this rearrangement is composed of at least two pericentric inversions that span ~100 Mb (Thomas et al. 2008; Davis et al. 2011). By utilizing limited loci from multiple chromosomes, Huynh et al. (2010) found reduced nucleotide diversity of both ZAL2 and ZAL2^m within the inversion relative to 1) a ~5 Mb region outside the rearrangement that still undergoes active recombination between the two chromosomes and to 2) other macrochromosomes in this species (Huynh et al. 2010). Huynh et al. (2011) and Davis et al. (2011) observed high differentiation between the two chromosomes and distinct ZAL2 and ZAL2^m haplotype clusters, supporting massively suppressed recombination (Davis et al. 2011; Huynh et al. 2011). These studies detected no evidence for genetic degeneration because of 1) no truncation of protein-coding genes, 2) no significant differences in non-synonymous substitution rates between the two chromosomes, and 3) little accumulation of repetitive sequences on ZAL2^m. Nevertheless, with data from WGS of multiple tan and white birds, Tuttle et al. (2016) observed a significant excess of non-synonymous polymorphisms on the ZAL2^m chromosome (Tuttle et al. 2016), conflicting with prior studies in this respect. Unlike the non-recombining W chromosome in which the majority of genes have decayed, ZAL2^m is similar in size to the ZAL2 chromosome, based on an

early karyotype study (Thorneycroft 1975); in addition, ZAL2^m homozygotes (or 'super-white'; ZAL2^m/ZAL2^m) are present in the population at ~0.2% frequency (Thorneycroft 1975; Horton et al. 2013; Tuttle et al. 2016), leading to the speculation that ZAL2^m may undergo a certain level of recombination. Consistent with this idea, a previous population genetics study predicted recombination events within the ZAL2^m chromosomes (Huynh et al. 2011).

In this study, we seek to close the following gaps. First, the degree of divergence between ZAL2 and ZAL2^m and of degeneration of ZAL2^m at the chromosome-wide scale is yet to be characterized. Second, a previous RNA-seq analysis identified co-expression networks linked to some behaviors at the morph level (Zinzow-Kramer et al. 2015), but the extent of regulatory divergence between ZAL2 and ZAL2^m alleles is unclear. Last, although the autosomal ZAL2/ZAL2^m system mimics sex chromosome systems regarding suppressed recombination and degeneration, we do not know whether dosage compensation, another feature of heteromorphic sex chromosomes, has evolved in response to disrupted expression of ZAL2^m-linked loci.

3.2 Results

3.2.1 Characterization of divergence between ZAL2 and ZAL2^m

As the reference genome of the white-throated sparrow (a tan bird, homozygous for ZAL2) is relatively fragmented (6,018 scaffolds), we used multiple approaches to identify scaffolds that are likely to reside on the ZAL2 chromosome. First, based upon well-maintained inter-chromosomal homology in birds (Shetty et al. 1999; Warren et al. 2010; Laine et al. 2016), we mapped scaffolds from the reference genome to the zebra finch genome and extracted those aligned to TGU3 (the zebra finch chromosome homologous to ZAL2/ZAL2^m). Second, we incorporated markers confirmed to be on ZAL2/ZAL2^m from

the previous fluorescent *in situ* hybridization (FISH) studies (Thomas et al. 2008; Davis et al. 2011; Tuttle et al. 2016). Third, we utilized homology with other passerine birds with chromosome-level assemblies collared flycatcher and great tit) (Ellegren et al. 2012; Laine et al. 2016) as additional evidence to support this designation.

Next, we distinguished the ZAL2 scaffolds inside or outside the rearrangement on the basis of levels of divergence. That is, non-recombining regions (inside the rearrangement) are expected to harbor higher differentiation relative to recombining regions (outside the rearrangement). For this aim, we calculated single nucleotide divergence (d_{XY}) and population differentiation (F_{ST}) (Weir and Cockerham 1984) between the tan and white birds using newly obtained whole-genome sequencing data of a rare super-white bird ($ZAL2^m/ZAL2^m$) and RNA-seq data from ~20 birds (Zinzow-Kramer et al. 2015) (Methods). Using these data, we also identified putatively fixed single nucleotide polymorphisms (SNPs) between ZAL2 and $ZAL2^m$ (Methods). Scaffolds with high d_{XY}, high F_{ST}, and a presence of potentially fixed SNPs are likely to be inside the rearrangement. In contrast, scaffolds with low d_{XY}, low F_{ST} and an absence of putatively fixed SNPs are likely to be outside the rearrangement.

Using the above methods, we identified 35 ZAL2 scaffolds (105.96Mb) that were likely inside the rearrangement, and five scaffolds (5.38Mb) that were outside. The regions inside and outside the rearrangement displayed significantly different levels of genetic differentiation. Inside the rearrangement, the average d_{XY} was 1.202 ± 0.006% (Figure 3.2A-B). On the other hand, the average d_{XY} outside the inversion showed variability that was statistically indistinguishable from that of the intra-species polymorphism (Figure 3.2A-B). Similarly, population differentiation (measured by F_{ST}) and insertion-deletion (indel) frequencies within the rearranged region were significantly higher than regions

outside it or in the rest of the genome (Figure 3.2C-D). These results were consistent with high genetic differentiation between ZAL2 and ZAL2^{m} inside the rearrangement.

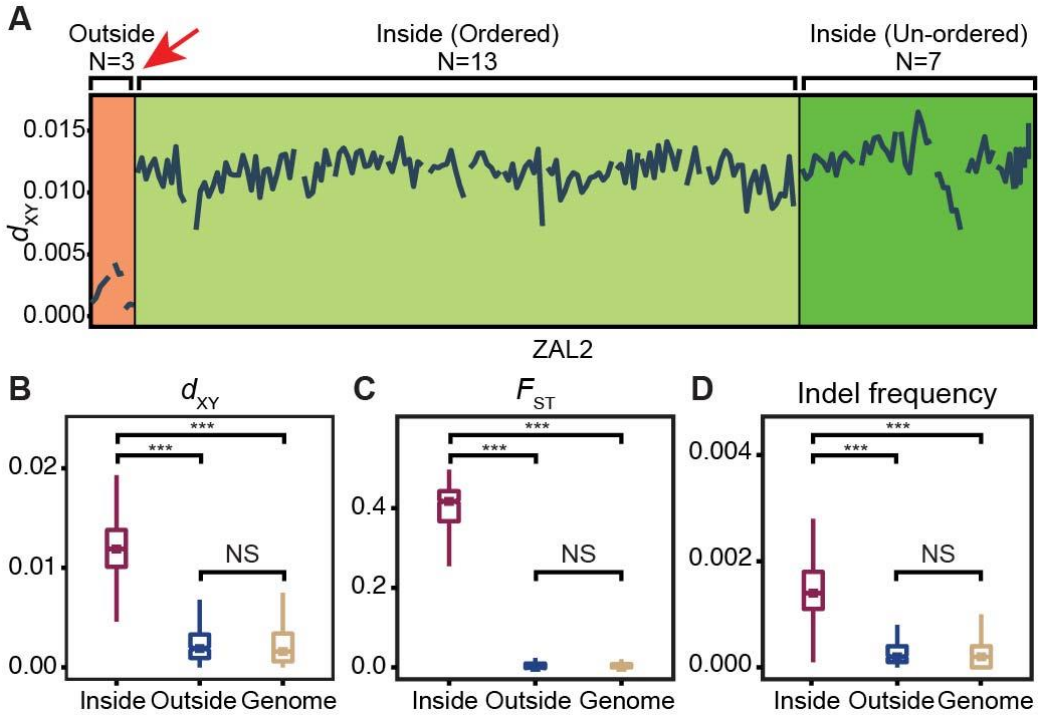

Figure 3.2 Genetic divergence between the ZAL2 and ZAL2^{m} chromosomes. (A) Pairwise nucleotide divergences (dXY) per 500 kb windows between the ZAL2 and ZAL2m chromosomes are shown for scaffolds > 1 Mb. The inferred major breakpoint (Thomas et al. 2008) is indicated with a red arrow. The scaffolds inside the rearrangement with precise locations known from previous fluorescent in situ hybridization (Thomas et al. 2008; Tuttle et al. 2016) were marked as "Ordered", and the others without known locations were marked as "Unordered". **(B-D)** Pairwise nucleotide divergences (d_{XY}), degrees of population differentiation (F_{ST}), indel frequencies, all measured in 10 Kb non-overlapping windows, were significantly higher in scaffolds within the rearrangement relative to those outside the rearrangement (Paired Mann–Whitney *U* test, NS: not significant; ***:*P* < 0.001).

3.2.2 Incipient degeneration of ZAL2^m-linked genes

Several genomic features of the ZAL2^m hint at its degeneration. Rates of non-synonymous substitution (d_N), rates of promoter (defined as 1kb upstream of transcription start site) substitution (d_P), the ratio of non-synonymous to synonymous substitution rates (d_N/d_S) as well as the ratio of radical to conservative amino acid substitution rates (d_R/d_C) were all elevated on ZAL2^m relative to ZAL2 (Figure 3.3). Nevertheless, sites that are mostly under neutral evolution [measured by rates of synonymous substitution (d_S) and whole-genome divergence (d_{XY})] did not show significant bias towards either chromosome (Figure 3.3). Increased substitutions at functional but not at neutral sites on the ZAL2^m chromosome suggested the accumulation of slightly deleterious mutations on ZAL2^m. Nevertheless, out of 1,007 protein-coding loci linked to the rearranged part of ZAL2^m, only 28 (~2%) contained premature stop codons or had lost start or stop codons. We found no frame-shift mutations or deletions of genes. Thus, the level of genetic degeneration of ZAL2^m is subtle.

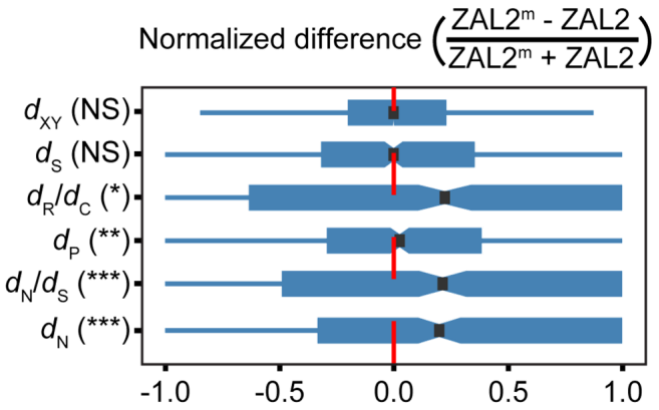

Figure 3.3 The degeneration of the ZAL2^m chromosome. The chromosomal bias of substitutions differs between putatively functional versus neutral sites. The normalized difference in substitution rates between ZAL2 and ZAL2^m is shown. This value is expected to be zero when there are equal numbers of ZAL2 and ZAL2^m substitutions (i.e., no bias), but will increase as the substitutions become more ZAL2^m-biased. Deviation from zero was tested using the Mann–Whitney *U* test and statistical significance is designated as NS: not significant; *:$P < 0.05$; **:$P < 0.01$; ***:$P < 0.001$. The measures presented include nucleotide divergence (d_{XY}), synonymous substitution rate (d_S), the ratio of radical to conservative amino acid substitution rates (d_N/d_S), promoter (defined as 1 kb upstream of transcription start site) substitution rate (d_P), the ratio of non-synonymous to synonymous substitution rates (d_N/d_S), and non-synonymous substitution rate (d_N), presented in the order of decreasing P values.

3.2.3 Substantial regulatory divergence between ZAL2 and ZAL2^m

Low levels of inter-chromosomal divergence and ZAL2^m degeneration would predict weak regulatory divergence between ZAL2 and ZAL2^m. To test that, we used RNA-seq datasets of hypothalamus and nucleus taeniae (two brain regions involved in social behaviors) samples from nine tan and ten white individuals (Zinzow-Kramer et al. 2015) to quantify allele-specific gene expression (ASE). Specifically, to minimize potential mapping bias towards the reference tan (ZAL2/ZAL2) genome caused by mismatches between ZAL2 and ZAL2^m, RNA-seq reads were aligned to an *N*-masked reference genome in which all fixed differences were masked and assigned to either chromosome.

Then, reads were assigned to either chromosome, expression levels were quantified, and differential expression was tested (see Methods for more details).

After this process, we found substantial expression divergence between ZAL2 and ZAL2^{m}. Specifically, 41.41% (335 genes) and 46.53% (375 genes) of all ZAL2/ZAL2^{m}-linked genes exhibited allele-specific expression in the hypothalamus and nucleus taeniae, respectively (Table 3.1). The two brain regions overlapped significantly in ASE genes (Figure 3.4), suggesting that genetic divergence of *cis*-regulatory regions may account for the observed expression divergence. As expected from ZAL2^{m} degeneration, we found a trend toward reduced ZAL2^{m} expression (one-tailed, paired Mann–Whitney *U* test, hypothalamus: $P = 0.066$ and nucleus taeniae: $P = 0.049$).

Table 3.1 ASE genes and their expression patterns between morphs.
Below, morph-biased genes are divided into two categories (tan > white and white > tan). Genes that do not exhibit significant morph-biased expression are shown in the 'Other' category. Differences were assessed by DESeq2, and numbers of genes with FDR-corrected $Q < 0.05$ are shown.

		tan > white	white > tan	Other (tan ≈ white)
Hypothalamus	ZAL2>ZAL2^{m}	63	1	111
	ZAL2^{m}>ZAL2	0	109	51
Nucleus Taeniae	ZAL2>ZAL2^{m}	54	1	153
	ZAL2^{m}>ZAL2	0	99	68

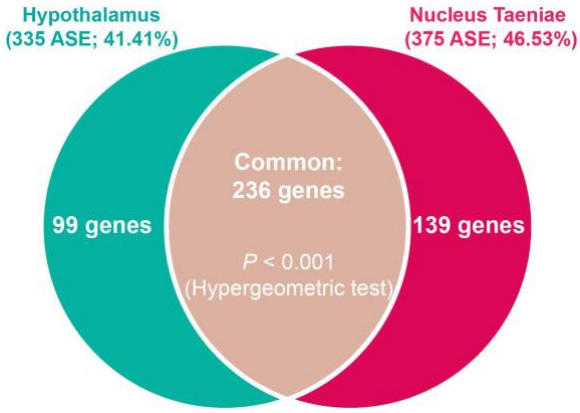

Figure 3.4 Significant overlap of ASE genes between ZAL2 and ZAL2^{m}.

Nevertheless, although the majority of ZAL2^{m} > ZAL2 genes (68%) also exhibited white > tan expression patterns, only a minority of ZAL2 > ZAL2^{m} genes (36%) were tan-biased (Figure 3.4; Figure 3.5). Therefore, we hypothesized that most ZAL2-biased genes, similar to genes on sex chromosomes across multiple lineages (Graves 2016), were dosage compensated.

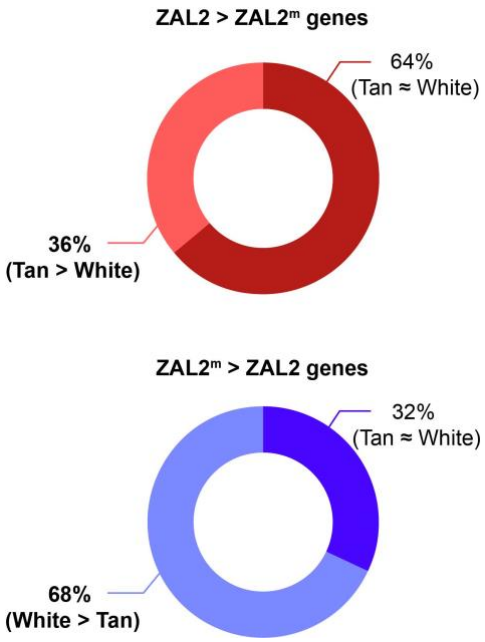

Figure 3.5 Pie charts to contrast the proportions of morph-biased genes for the two categories of ASE genes (ZAL2- and ZAL2^{m}-biased).

3.2.4 Potential dosage compensation in the autosomal system

To test dosage compensation in this autosomal system, we proposed the following model. Reduced expression in ZAL2^{m} caused by degeneration at regulatory regions could lead to lower total expression in the white birds relative to tan birds (Figure 3.6A-B). Natural selection may increase the expression of ZAL2 alleles in the white birds to bring their expression close to that of tan birds, in which expression reflects the ancestral and presumably optimal levels of gene expression (Figure 3.6C). In this scenario, ZAL2 allelic expression in white birds should be higher than that in tan birds (white-ZAL2 / tan-ZAL2 > 1).

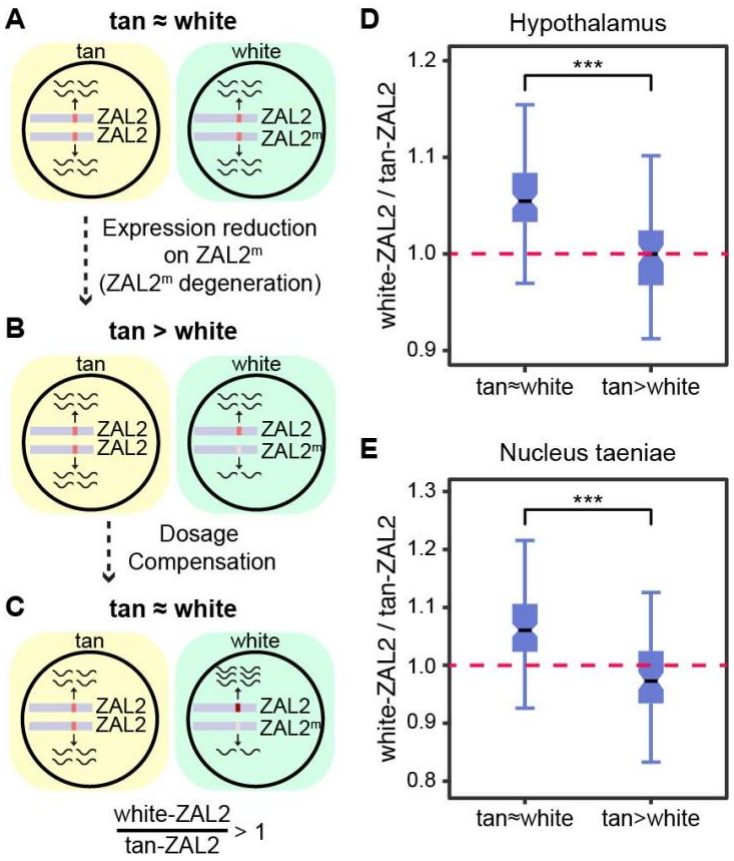

Figure 3.6 Potential dosage compensation for ZAL2-biased genes. (A) Before degeneration, expression dosage (black waves) is similar between ZAL2 and ZAL2^{m} and between tan and white. **(B)** If the expression of ZAL2^{m} alleles is reduced and there is no dosage compensation, white individuals should show reduced expression. **(C)** The dosage between the morphs may be re-balanced via overexpression of the ZAL2 allele in white birds. Consequently, the expression of the ZAL2 allele should be greater in white than tan birds (white-ZAL2 / tan-ZAL2 > 1). **(D-E)** Levels of compensation (measured by white-ZAL2 / tan-ZAL2) are significantly elevated for tan ≈ white (dosage-compensated) genes compared with non-dosage compensated genes (tan > white) (Mann–Whitney *U* test, ***:*P* < 0.001).

For ZAL2-biased genes in our datasets, we observed the expected patterns in both brain regions. Specifically, tan ≈ white genes (dosage-rebalanced genes) displayed higher expression of ZAL2 alleles in white relative to tan birds, but tan > white genes (dosage-unbalanced genes) did not (Figure 3.6D-E). Intriguingly, for ZAL2^{m}-biased genes that were

not morph-biased, we also observed altered ZAL2 expression in white relative to tan birds (tan-ZAL2 / white-ZAL2 > 1) (**Figure B. 1**), suggesting that degeneration in the regulatory regions of ZAL2^{m}-linked genes could sometimes lead to overexpression. Taken together, our results support potential dosage compensation in the ZAL2/ZAL2^{m} system.

3.2.5 ASE genes that are not dosage compensated exhibit higher connectivity in co-expression networks

We hypothesized that ASE genes that were not dosage compensated may play essential roles in mediating morph differences. To test this hypothesis, we examined previously identified weighted co-expression networks from our RNA-seq data (Zinzow-Kramer et al. 2015). Genes in two large modules, namely the 'black' module in the hypothalamus (containing 511 genes, 226 genes on ZAL2/ZAL2^{m}) and the 'greenyellow' module in the nucleus taeniae (containing 157 genes, 115 genes on ZAL2/ZAL2^{m}), were significantly correlated with territorial singing, a behavior that differs between the morphs (Zinzow-Kramer et al. 2015). Consistent with our hypothesis, ASE genes exhibited significantly higher intramodule connectivity compared with other genes in the modules, indicating their central roles in co-expression networks (Figure 3.7).

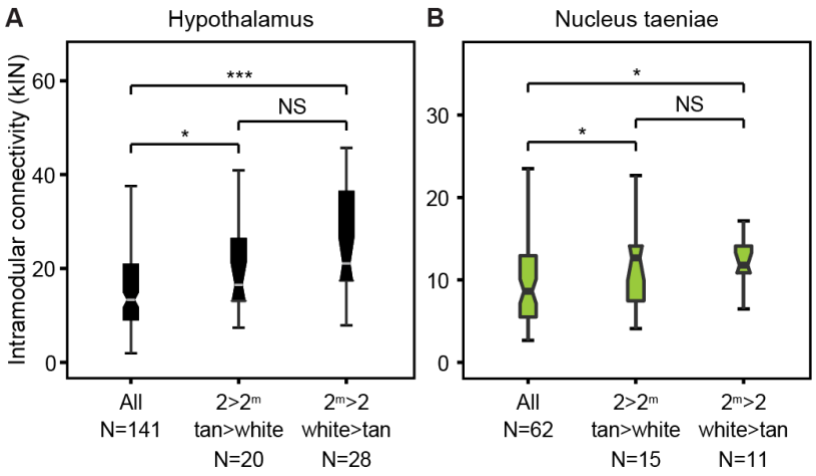

Figure 3.7 Allele- and morph-biased genes occupy more central positions in gene co-expression networks. Intramodular connectivity (kIN) for all genes in modules compared with those exhibiting allele-specific expression for **(A)** hypothalamus, and **(B)** nucleus taeniae. Mann-Whitney *U* test *:$P < 0.05$; ***: $P < 0.001$; NS: $P > 0.05$.

3.3 Discussion

The ZAL2/ZAL2^m system in the white-throated sparrow is linked to two plumage morphs with behaviors complementing each other (Thorneycroft 1966; Tuttle 2003; Maney et al. 2015). Same-morph mating is largely restricted. Based on several reports (Thorneycroft 1975; Michopoulos et al. 2007; Falls and Kopachena 2010; Horton et al. 2013; Tuttle et al. 2016), only 0.8% of total matings occur between white birds, and as a result, the frequency of ZAL2^m/ZAL2^m homozygotes is ~0.2% of total offspring.

Owing to this mating strategy, the recombination of ZAL2^m is largely limited. To characterize the divergence between ZAL2 and ZAL2^m, we sequenced the genome of a rare super-white bird homozygous for ZAL2^m. Comparative analysis with the new data revealed a divergence of ~1.2%, comparable to that between humans and chimpanzees

(e.g., (Elango et al. 2006)). Similar to sex chromosomes in multiple lineages, the recombination-suppressed ZAL2^m chromosome displays several signs of degeneration at functional sites. Nevertheless, the degree of ZAL2^m degeneration is weak, evidenced by very few genes undergoing pseudogenization. Indeed, individuals homozygous for the ZAL2^m chromosomes are found in nature at low frequencies (Thorneycroft 1975; Michopoulos et al. 2007; Falls and Kopachena 2010; Horton et al. 2013; Tuttle et al. 2016), indicating that ZAL2^m is functional.

Although we found only low levels of genetic divergence between ZAL2 and ZAL2^m as well as weak degeneration, we found strong signals of allele-specific expression for more than 40% of genes in white birds. Many of these genes showed reduced expression of the ZAL2^m allele, which could indicate degeneration of regulatory sequences on the ZAL2^m chromosome. Our results are consistent with previous findings that regulatory divergence can precede large-scale protein sequence differentiation (King and Wilson 1975; Carroll 2005; Zhou and Bachtrog 2012).

Disrupted expression of genes on the non-recombining chromosome can initiate the evolution of dosage compensation, re-balancing expression between heterozygotes and homozygotes. A variety of such mechanisms has evolved in sex chromosomes (Meyer 2010; Conrad and Akhtar 2011; Graves 2016), and it is known that dosage compensation can evolve rapidly in new sex chromosomes (e.g., (Muyle et al. 2012; Papadopulos et al. 2015)). However, to our knowledge, dosage compensation has not been reported in autosomal systems. Here, we show that for a number of ASE genes, dosage compensation may exist to re-balance expression between morphs. If we define the degree of dosage compensation as the ratio of the expression of the compensated allele to the expression of the non-compensated allele (for example, white-ZAL2 relative to tan-ZAL2 in the case of a reduction in white-ZAL2^m), the average degree of dosage

compensation is approximately ~35% in both of the brain regions we looked at. This level is similar to that for avian sex chromosomes; Z-linked expression in females (ZW) is only 30-40% elevated compared with that in males (ZZ) (Ellegren et al. 2007; Itoh et al. 2007; Uebbing et al. 2013; Graves 2016). It is of interest to understand whether some mechanisms for dosage compensation are shared between the ZAL2/ZAL2^{m} and Z/W systems.

Taken together, our comprehensive genetic and transcriptomic study in the white-throated sparrow reveals that protein-coding sequences and gene expression are evolving with strikingly different patterns. Whereas protein-coding sequences harbor few genetic differences, we found strong evidence for the rapid evolution of regulatory divergence as well as dosage compensation.

3.4 Methods

3.4.1 Sequencing of a super-white bird

We sequenced a super-white bird homozygous for the rearrangement (ZAL2^{m}/ZAL2^{m}) (Horton et al. 2013). High molecular weight genomic DNA was extracted from liver and sequenced using HiSeq2500 at the Roy Carver Genome Center of the University of Illinois. Approximately 240 million 150 bp reads were generated.

3.4.2 Identification of scaffolds on the second chromosome

Genomic scaffolds from a tan bird (ZAL2/ZAL2) were recently published by Tuttle et al. (2016). To confidently identify scaffolds that originate from the second chromosome of white-throated sparrows, we mapped those scaffolds onto that of the zebra finch using LASTZ 1.03.73 (Harris 2007) (parameters: --step=20 --chain --gfextend --gapped --

traceback=2000M --ydrop=300400 --identity=85 --matchcount=1000), with additional parameters of >30% coverage for scaffolds longer than 10 kb and >80% coverage for scaffolds shorter than 10 kb. These cutoff criteria were selected to avoid scaffolds that mapped to ZAL2 due to partial mapping to repetitive sequences (Zhou et al. 2014), on the basis of the coverage% distribution across scaffolds (**Figure B. 2**). We identified ZAL2 scaffolds using homology to the corresponding zebra finch chromosome (commonly referred to as TGU3 due to its homology to chicken chromosome 3 (Warren et al. 2010)), previous fluorescent *in situ* hybridization (FISH) studies (Thomas et al. 2008; Davis et al. 2011; Tuttle et al. 2016), as well as homology with two other passerine birds with chromosome-level assemblies, collared flycatcher (*Ficedula albicollis*) and great tit (*Parus major*) (Ellegren et al. 2012; Laine et al. 2016). Following these procedures, we identified 56 scaffolds on the ZAL2 chromosome.

Compared with the results of Tuttle et al. (Tuttle et al. 2016), our results included 19 additional ZAL2 scaffolds (corresponding to ~25Mb) that were previously unrecognized as linked to this chromosome. In addition, we excluded a ~45Mb scaffold (NW_005081536.1) that had been denoted as residing on a non-rearranged portion of the ZAL2 (Tuttle et al. 2016). Regions homologous to this scaffold were found on a different chromosome in the zebra finch, collared flycatcher, and great tit, making it unlikely that it has moved to ZAL2 in *Z. albicollis* given the well-conserved chromosomal homology in birds (Shetty et al. 1999; Warren et al. 2010; Laine et al. 2016). FISH studies also previously showed that regions outside of the rearrangement on chromosome two are only ~10Mbs in length (Thomas et al. 2008; Davis et al. 2011).

3.4.3 Genetic divergence between ZAL2 and ZAL2^m from genome sequences and RNA-seq data

We then called variants (SNPs and indels) that distinguish ZAL2 and ZAL2^m sequences by following the Genome Analysis Toolkit (GATK) best practices for variant calling in genome sequencing data (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). First, super-white reads from whole-genome sequencing were aligned to the reference genome from a tan bird (Tuttle et al. 2016) using BWA 0.7.12 (Li and Durbin 2010). GATK 3.4 was used to call variants, and those variants with quality lower than 30 or read depth less than 5 were excluded. We then used a sliding window approach to calculate short indel frequencies and d_{XY} following Jukes-Cantor correction (Jukes and Cantor 1969).

We also called variants from available transcriptome data from the hypothalamus and nucleus taeniae of nine tan and 11 white individuals (Zinzow-Kramer et al. 2015) following the GATK best practices for variant calling in RNA-seq data (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). A variant was considered putatively fixed in the sampled individuals if: 1) the variant was biallelic; 2) tan individuals were homozygous for the reference allele (AA); 3) white individuals (ZAL2/ZAL2^m) were heterozygous (Aa); 4) the super-white (ZAL2^m/ZAL2^m) individual was homozygous for the alternative allele (aa).

3.4.4 Scaffolds inside versus outside the rearrangement.

Scaffolds residing inside versus outside the rearrangement were identified on the basis of distinctive patterns of genetic divergence. Specifically, the distributions of d_{XY} and F_{ST} were bimodal, categorized as 'high' or 'low' d_{XY} or F_{ST} (**Figure B. 3**). Scaffolds with high d_{XY}, high F_{ST}, and putatively fixed differences were designated as 'confidently inside'. If only one of the two criteria were satisfied, the scaffolds were designated as 'likely inside'. Scaffolds were defined as 'confidently outside' if they had low d_{XY}, low $F_{ST,}$ and no

putatively fixed differences, and one of the following two conditions was satisfied: either they were already shown to be outside the rearrangement (Thomas et al. 2008; Davis et al. 2011; Tuttle et al. 2016), or they shared homology with the homologous chromosome of two other passerine birds (*F. albicollis* and *P. major*) (Ellegren et al. 2012; Laine et al. 2016). Scaffolds that exhibited low d_{XY}, low F_{ST} and an absence of putatively fixed differences, with no extra supporting evidence, were designated as 'likely outside'.

3.4.5 Analyses of protein-coding sequences.

For each ZAL2-linked gene, we extracted the longest transcript and constructed the ZAL2^m version with the putatively fixed differences. Additionally, we downloaded all available genome annotations for 13 other avian species in the order of Passeriformes (the same order as the white-throated sparrow) from NCBI, and the tree for all 14 species was inferred from several avian phylogeny studies (Jetz et al. 2012; Jarvis et al. 2014; Jetz et al. 2014; Prum et al. 2015). Each gene was aligned by MAFFT v7.245 (Katoh and Standley 2013), low-quality alignment parts were trimmed by trimAl v1.4 (Capella-Gutierrez et al. 2009), and the codon alignment was constructed by PAL2NAL v14 (Suyama et al. 2006). We obtained codon alignments for a total of 800 genes. We calculated d_N and d_S for the ZAL2 and ZAL2^m branches with a free-ratio model using codeml from the PAML 4.8 package (Yang 2007). The Hon-New package (Zhang 2000), which adopts the amino acid classification system that considers charge and polarity, was used to estimate rates of radical amino acid substitution (d_R) and conservative amino acid substitution (d_C).

3.4.6 Gene expression.

We examined genes with morph-biased (Tan≠White) expression and allele-specific expression (ZAL2≠ZAL2^{m}) patterns. To account for potential mapping bias towards the reference (ZAL2/ZAL2) genome caused by mismatches between ZAL2 and ZAL2^{m}, we *N*-masked (Krueger and Andrews 2016) putatively fixed differences in the reference. We additionally checked potential left-over bias by aligning whole-genome sequences from three white birds described by Tuttle et al. (Tuttle et al. 2016) to this *N*-masked genome. ZAL2 and ZAL2^{m} alleles should have roughly equal coverage per site if mapping bias has been eliminated. Indeed, for all three white birds, we did not observe significant coverage bias towards the ZAL2 allele (see '*Examining mapping bias in the N-masked reference genome*' and **Figure B. 5**).

We mapped the aforementioned RNA-seq data from nine tan and 10 white individuals (Zinzow-Kramer et al. 2015) to the *N*-masked genome with STAR 2.4.1d under the 2-pass mode (Dobin et al. 2013). Only uniquely mapped reads were retained for further differential expression analysis. SNPsplit 0.3.3 (Krueger and Andrews 2016) was run to assign reads to ZAL2 or ZAL2^{m} for the white samples and to filter out reads without fixed differences in the tan samples. Read counts per gene at the morph and allele level were calculated by htseq-count 0.9.1 (Anders et al. 2015) with '-s no -m intersection-nonempty'.

To detect morph-biased expression, we calculated size factors, normalized libraries with these factors, and then identified differential expression with 'design = ~ morph' in DESeq2 1.12.3 (Love et al. 2014). To detect allele-specific expression, we normalized libraries with the size factors generated in the previous step and identified differential expression with 'design = ~ allele + sample' in DESeq2. Only genes with average expression counts ('baseMean' in the DESeq2 output) higher than 5 at the morph level were retained for later analysis (809 genes for the hypothalamus and 806 for nucleus taeniae).

3.4.7 De novo assembly of the super-white genome, whole-genome alignment, and detection of gene deletion

Paired-end sequences from the super-white bird were first trimmed by PRINSEQ 0.20.4 (Schmieder and Edwards 2011) and assembled by Abyss 1.5.2 (Simpson et al. 2009). The final assembly has a contig N50 of 26,601bp and a scaffold N50 of 32,876 bp. The total assembly size is 1.01 Gb, which is close to the estimated genome size of the white-throated sparrow (Tuttle et al. 2016). We aligned the newly generated super-white assembly to the tan reference genome using LASTZ with the parameters mentioned above. We found no evidence of deletion of exons and/or large (> 50 bp) indels.

3.4.8 Examining mapping bias in the N-masked reference genome.

We examined potential leftover mapping bias by aligning whole-genome sequences of three white birds (Sample IDs: 10_083, 10_092 and 10_093) published by Tuttle et al. (Tuttle et al. 2016) to the N-masked genome using HISAT2.1.0 (Kim et al. 2015) (parameters: --no-spliced-alignment --sp 1000,1000). Reads were assigned to the ZAL2 or ZAL2^{m} chromosome by SNPsplit 0.3.3 (Krueger and Andrews 2016) on the basis of putative fixed differences, and bedtools genomecov (Quinlan and Hall 2010) was used to count per base coverage on ZAL2 and ZAL2^{m}, respectively. Across putative fixed differences, we observed roughly equal per base coverage between ZAL2 and ZAL2^{m} (**Figure B. 5**), suggesting mapping bias was significantly eliminated using the SNP *N*-masking approach.

3.5 Acknowledgements

This chapter has been published in *Proceedings of the National Academy of Sciences of the United States of America* (Sun et al. 2018). Based on the journal's rights

and permission policy (https://www.pnas.org/page/about/rights-permissions), coauthors and I retain copyright to this work, and it can be included in my dissertation without further need to obtain permissions.

This work is a collaboration between Dr. Soojin Yi's lab at Georgia Tech and Dr. Donna Maney's lab at Emory University. We thank the Roy Carver Genome Center of the University of Illinois for help with sequencing; Dr. Kathleen Grogan, members of the Yi laboratory, Dr. Brian Charlesworth, and two anonymous reviewers for comments on the manuscript; and Xiulan Pan and Dr. Chen Feng for help with illustration. This work was supported by National Institutes of Health Grants R01 MH082833 (to D.L.M.) and 1R21 NIMH102677 (to D.L.M. and S.V.Y.) and by Georgia Tech (S.V.Y.).

CHAPTER 4

EPIGENETIC DIFFERENTIATION BETWEEN NON-RECOMBINING AUTOSOMES IN THE WHITE-THROATED SPARROW

4.1 Introduction

The genetic differentiation of the second chromosome in the white-throated sparrow (*Zonotrichia albicollis*) is completely linked to two distinct plumage morphs, tan and white. One distinguishing feature of the two morphs is their behaviors: white birds (ZAL2/ZAL2^{m}) engage in more territorial aggression but invest less in parenting behavior than tan birds (ZAL2/ZAL2) (Tuttle 2003; Maney 2008; Horton et al. 2014a; Horton et al. 2014b; Maney et al. 2015; Zinzow-Kramer et al. 2015). Birds of the two morphs mate in a highly disassortative manner (Tuttle et al. 2016), causing strongly suppressed recombination between ZAL2 and ZAL2^{m} chromosomes. Despite a low level of nucleotide divergence (~1%) between the two chromosomes, more than 40% of genes are differentially expressed in the brain, suggesting rapid regulatory evolution which precedes large-scale genetic differences in the ZAL2/ZAL2^{m} system (Sun et al. 2018). However, the molecular basis of such differential gene expression is not resolved. In this chapter, we investigated DNA methylation as a potential regulatory driver of differences in expression.

The major form of DNA methylation is the addition of methyl groups to the 5th carbon of cytosines (5mC) (Bird 1992; Bird 2002). In mammalian systems, studies have established a model in which methylation at CpG sites of *cis*-regulatory regions can lead to transcription repression (Schübeler 2015). Because of its regulatory role, DNA methylation is essential to numerous biological processes, such as genomic imprinting, sexual differentiation, X chromosome inactivation, and maintaining genome stability

(Robertson and A.Jones 2000; Melamed and Arnold 2007; Roeszler et al. 2012; Graves 2016). DNA methylation has also been implicated in development and aging. For example, studies have demonstrated that methylation of “clock CpGs” (CpGs that change DNA methylation with age) is a relatively accurate predictor of chronological age (Horvath 2013; Petkovich et al. 2017). Furthermore, a previous study identified “aging segments” (clusters of CpGs that respond to age) and showed that segments that increase DNA methylation are enriched in genes associated with development (Sun and Yi 2015).

In this study, using deep whole-genome bisulfite sequencing (WGBS), we generated single-nucleotide-resolution methylation maps of 12 female brain samples from both morphs and from two developmental time points, namely seven adult (four white and three tan) and five chick (three white and two tan) samples. To connect DNA methylation to gene expression, we integrated RNA sequencing (RNA-seq) data from the same samples with these methylomes. Our comprehensive epigenetic study in the white-throated sparrow allows us to understand how variations in developmental stages and plumage morphs are associated with the methylation landscapes of this species.

4.2 Results

4.2.1 Genome-wide characterizations of DNA methylation

The 12 individuals analyzed here are not related, inferred by a kinship analysis using their SNP data (**Figure C. 1**) (Manichaikul et al. 2010). A principal component analysis (PCA) revealed strong age effects on DNA methylation (Figure 4.1A). That is, the first principal component (PC1), which distinguished adults from chicks, accounted for the largest amount of variation in DNA methylation (~20%), and the PC2 explained ~11% sample variation by morph. When considering only the CpGs within the rearrangement

ZAL2/ZAL2^{m} chromosome, samples were separated by both age and morph even clearer (Figure 4.1B).

To detect significant differentially methylated CpGs (referred to as 'DMCs') between adults and chicks as well as between white and tan birds, we performed a differential methylation analysis for individual CpGs using DSS (Wu et al. 2015). Specifically, when identifying DMCs between adults and chicks (referred to as 'age-DMCs'), we treated age as the test variable and adjusted effects from morph; in contrast, when identifying DMCs between tan and white birds (referred to as 'morph-DMCs'), we treated morph as the test variable and adjusted effects from age. This analysis revealed 286,434 age-DMCs and 4,507 morph-DMCs (FDR < 0.05 and absolute difference in DNA methylation ≥ 10%). While age-DMCs were distributed across the genome (except for the Z chromosome), morph-DMCs were restricted to the ZAL2/ZAL2^{m} chromosomes (Figure 4.1C). This finding was consistent with age being the main contributor to methylation variation in the PCA analysis. Interestingly, although age-DMCs were widespread across the genome, they had smaller effect sizes (absolute differences in DNA methylation between two groups in comparison) compared with morph-DMCs (Figure 4.1D).

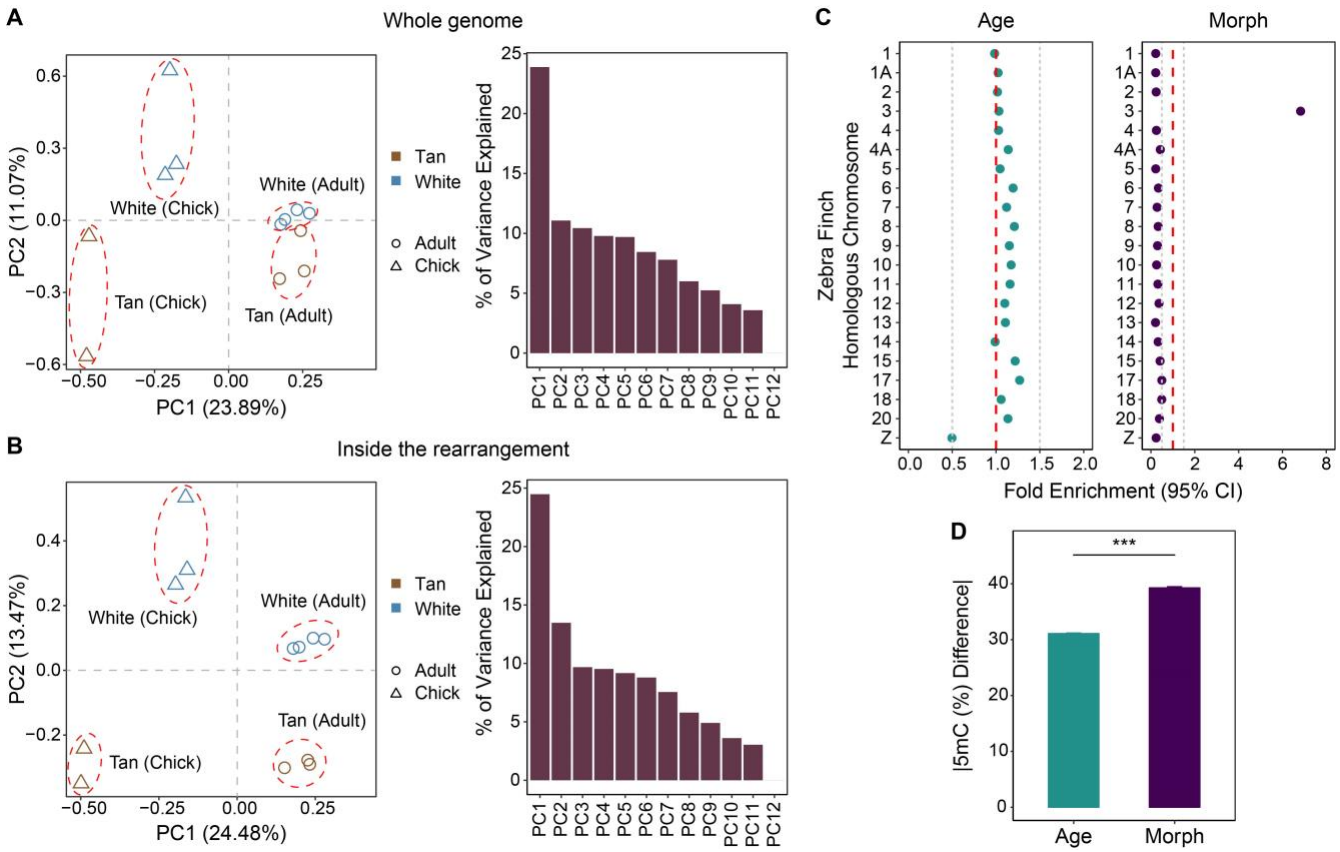

Figure 4.1 The effects of age and morph on DNA methylation patterns. PCA of WGBS samples for **(A)** all CpGs and **(B)** CpGs within the rearrangement. The scree plots displaying percent variance explained by each PC were drawn on the right. **(C)** Fold enrichment with 95% confidence interval (95% CI) for the chromosome distribution of DMCs (using homologous chromosomes in zebra finch for designation). Note that ZAL2 and ZALZ correspond to 3 and Z in the zebra finch, respectively. The fold enrichment and confidence intervals were calculated by comparing the real distribution of DMCs with the null distribution generated by 100 random selections of the same number of CpGs. The red dashed lines indicate no depletion/enrichment (1) of DMCs on a chromosome, and the gray dashed lines depict boundaries for moderate depletion (0.5) or enrichment (1.5) of DMCs. Only chromosomes larger than 10 Mb were shown. **(D)** Mean absolute differences in DNA methylation (5mC [%]) of age-DMCs and morph-DMCs. Relative to morph-DMCs, age-DMCs had smaller effect sizes. Significance was assessed with the Mann-Whitney *U* test (***:$P < 0.001$).

4.2.2 Global hypermethylation of CpGs in adults relative to chicks

Next, we asked whether age-related changes in DNA methylation were associated with DNMT expression. We found that age-DMCs were predominantly hypermethylated in

adult relative to chick samples (Figure 4.2A). Accordingly, DNA methyltransferases DNMT1 and DNMT3b had significantly higher expression in adults than in chicks (Figure 4.2B; note that DNMT3a is not annotated in the reference genome due to the poor assembly quality around that region). Interestingly, age-DMCs located in promoters were significantly enriched for gene ontology (GO) terms related to the development and cell differentiation (Figure 4.2C).

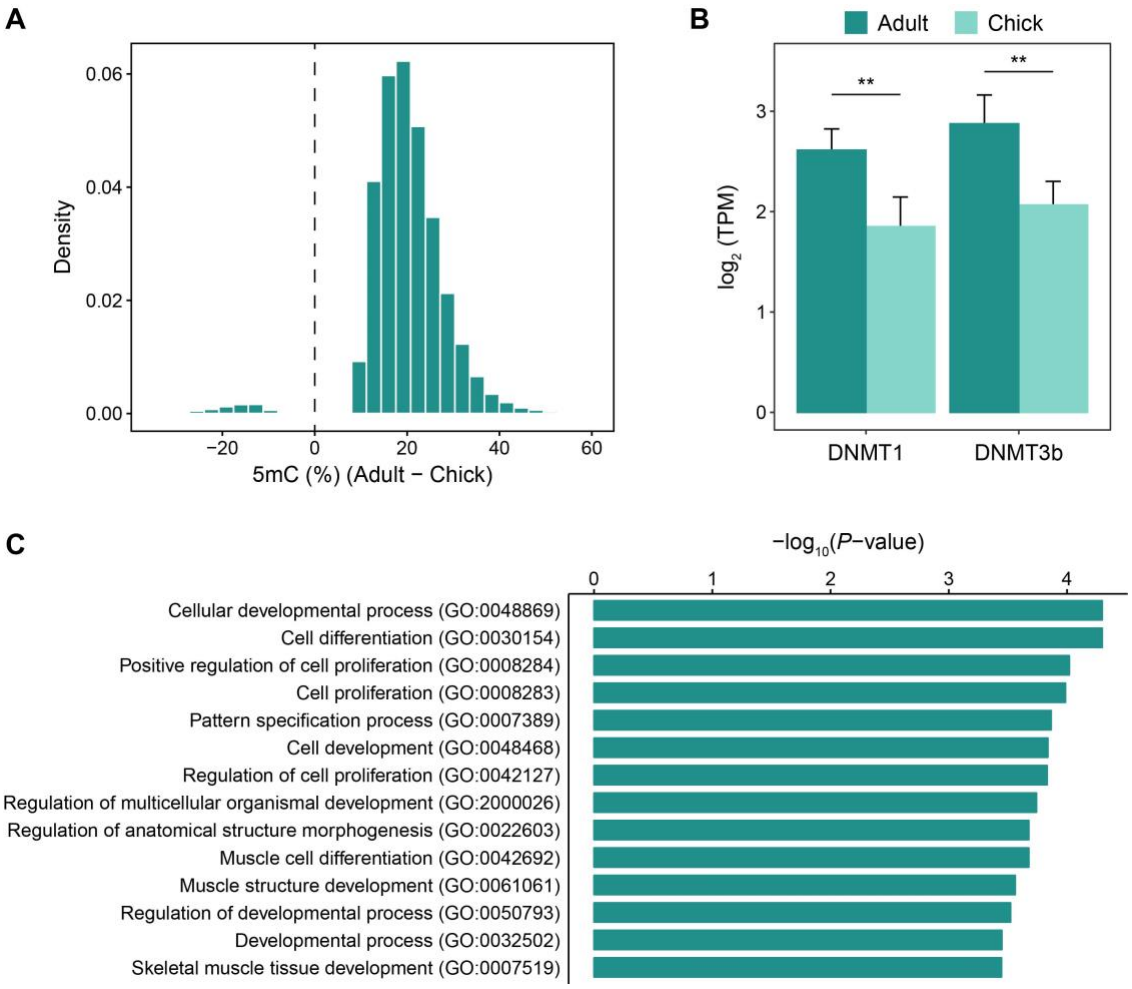

Figure 4.2 Hypermethylation in adults relative to chicks and its potential functional implications. **(A)** A histogram to show age differences in DNA methylation. The vast majority of age-DMCs were hypermethylated in adults compared with chicks. **(B)** Both DNMT1 and DNMT3b were overexpressed in adults (tested by DESeq2, ***:$P < 0.001$), consistent with hypermethylation in adults. **(C)** GO enrichment of genes that contain at least three age-DMCs within promoters (defined as within 1.5 Kb upstream of TSS). A statistical overrepresentation test was performed by PANTHER14.1 (Fisher's exact test), with all white-throated sparrow genes present in the *Gallus gallus* annotation database as the reference list. Only GO terms with $Q < 0.05$ and fold enrichment > 1.5 were reported.

4.2.3 Methylation landscapes of ZAL2 and ZAL2^m

As morph effects on DNA methylation were mostly exclusive to ZAL2/ZAL2^{m} chromosomes, we investigated DNA methylation patterns of the two chromosomes. By adjusting for age effects in the regression model using DSS, we detected 13,774 CpGs that were differentially methylated between ZAL2 and ZAL2^{m} (referred to as 'allele-DMCs'). The distribution of allelic differences in DNA methylation revealed three notable classes of allele-DMCs, with two groups of ZAL2^{m}-hypomethylated CpGs and a group of ZAL2^{m}-hypermethylated CpGs (Figure 4.3A). For further analysis, we used a 50% of allelic differences in DNA methylation to distinguish the two groups of ZAL2^{m}-hypomethylated CpGs and assigned all allele-DMCs to one of the three distinct classes: extremely ZAL2^{m}-hypomethylated (henceforth ZAL2^{m} << ZAL2; left peak), ZAL2^{m}-hypomethylated (henceforth ZAL2^{m} < ZAL2; middle peak) and ZAL2^{m}-hypermethylated (henceforth ZAL2^{m} > ZAL2; right peak) (Figure 4.3A).

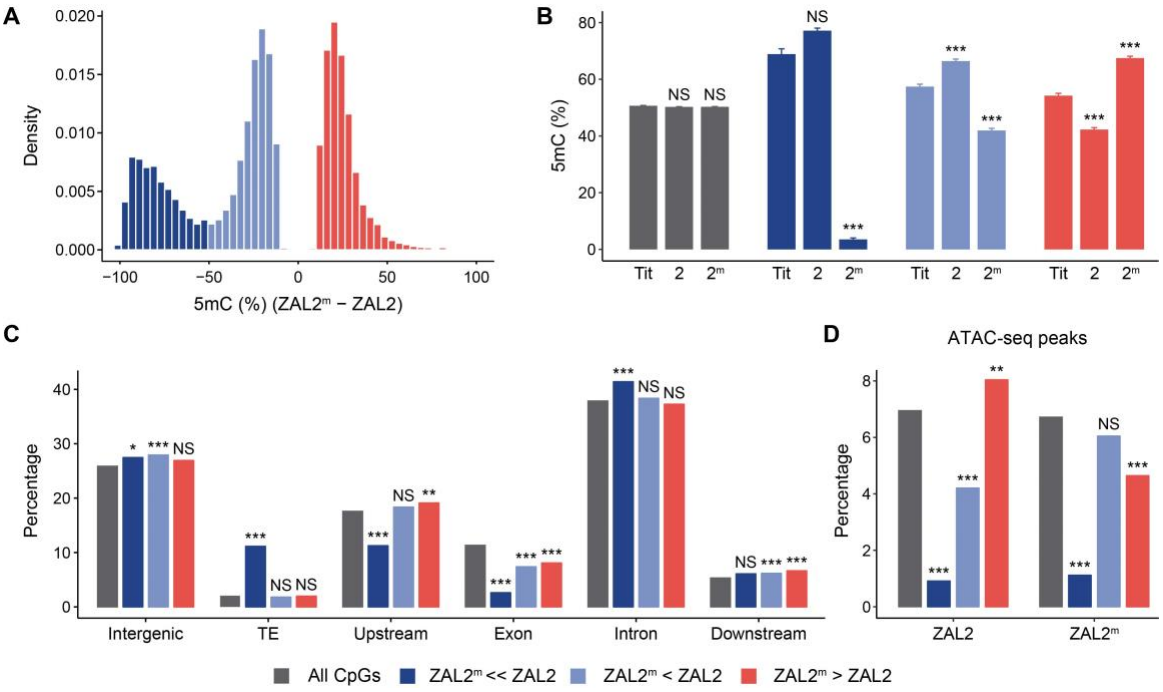

Figure 4.3 Characterizations of the three classes of allele-DMCs. (A) A histogram to show allelic differences in DNA methylation. **(B)** Changes in DNA methylation levels relative to the ancestral methylation levels inferred by an outgroup species (great tit). Mann–Whitney *U* test was used to assess the significance. **(C)** Percentage of allele-DMCs within different genomic regions. Intergenic regions were defined as regions that were at least 10 Kb away from any genes, and upstream/downstream distal regions were defined as 10 Kb upstream/downstream of the transcription start site (TSS)/transcription end site (TES). **(D)** Percentage of allele-DMCs within ZAL2 or ZAL2^{m} ATAC-seq peaks. For C-D, all ZAL2/ZAL2^{m}-linked CpGs were used as the control, and enrichment or depletion was assessed by two proportion Z-test. For B-D, NS: not significant; *:$P < 0.05$; **:$P < 0.01$; ***:$P < 0.001$.

Were the allelic differences in DNA methylation due to changes in DNA methylation on ZAL2 or changes in DNA methylation on ZAL2^{m} relative to the ancestral methylation levels? For this aim, we used an approach similar to what was described in Chapter 2 (Section 2.2.5), but for DNA methylation instead of gene expression. Specifically, for each class of DMCs, we compared methylation levels of conserved CpGs on ZAL2 or ZAL2^{m} to that of outgroup species with available methylome data for the brain (great tit, which is

also a passerine bird). Because the vast majority of ZAL2/ZAL2^{m}-linked CpGs were not differentially methylated between the two chromosomes, we hypothesized that DNA methylation levels of these CpGs were similar between the white-throated sparrow and great tit (representing ancestral methylation levels). Indeed, we did not detect significant differences in methylation levels between the two avian species (Figure 4.3B). In contrast, for DMCs, epigenetic alterations relative to ancestral levels appeared to occur on both alleles. Intriguingly, for ZAL2^{m} << ZAL2 DMCs, a massive ZAL2^{m} decrease but only a mild and nonsignificant ZAL2 increase in DNA methylation was detected (Figure 4.3B), suggesting that strong ZAL2^{m} hypomethylation was potentially the major driving force of the epigenetic evolution of this class of CpGs.

Next, we investigated whether the three classes of allele-DMCs exhibited distinct genomic features. We found that, for both classes of DMCs with smaller absolute allelic differences in methylation (ZAL2^{m} < ZAL2 and ZAL2^{m} > ZAL2), the distances between adjacent DMCs were significantly shorter than the null distribution generated by 100 random selections of the same number of CpGs from the ZAL2/ZAL2^{m} chromosome; conversely, the distances between ZAL2^{m} << ZAL2 DMCs were longer than the null distribution (**Figure C. 2**). This observation suggested that ZAL2^{m} extremely hypomethylated CpGs were less likely than the other two classes of DMCs to cluster into regions. In line with this idea, ZAL2^{m} << ZAL2 DMCs were significantly enriched in transposable elements (TEs) and intronic regions in which CpGs were generally sparsely distributed, and they were significantly depleted from regions upstream of TSS in which CpG islands were typically located (Figure 4.3C).

Using ATAC-seq data of a white sample from the same brain region (Methods), we found that ZAL2^{m} << ZAL2 DMCs were strongly depleted from regions with high chromatin accessibilities (mostly transcriptionally active regions) (Figure 4.3D).

Interestingly, ZAL2^{m} < ZAL2 DMCs were depleted from ZAL2 peaks but not in ZAL2^{m} peaks, and ZAL2^{m} > ZAL2 DMCs were enriched in ZAL2 peaks but depleted from ZAL2^{m} peaks (Figure 4.3D). This finding implied a potentially negative relationship between allelic differences in DNA methylation and allelic differences in gene expression, which we will explore in the next section.

4.2.4 How are allelic differences in methylation associated with allelic differences in expression?

To test the regulatory role of allele-DMCs, we performed a differential expression analysis between alleles using DESeq2 (Love et al. 2014) (Methods). Almost 75% of ASE genes were ZAL2-biased (243 ZAL2-biased vs. 87 ZAL2^{m}-biased genes; FDR-corrected *Q*-value < 0.05), supporting global silencing caused by ZAL2^{m} degeneration. For both promoter (defined as 1.5 Kb upstream of TSS) and gene body regions, allelic differences in DNA methylation and allelic differences in gene expression were negatively associated (Figure 4.4A). When ZAL2^{m} << ZAL2 DMCs were excluded from this analysis, the negative correlation became stronger in gene body regions (Figure 4.4B). No GO terms were enriched in genes containing allele-DMCs in their promoter or gene body regions.

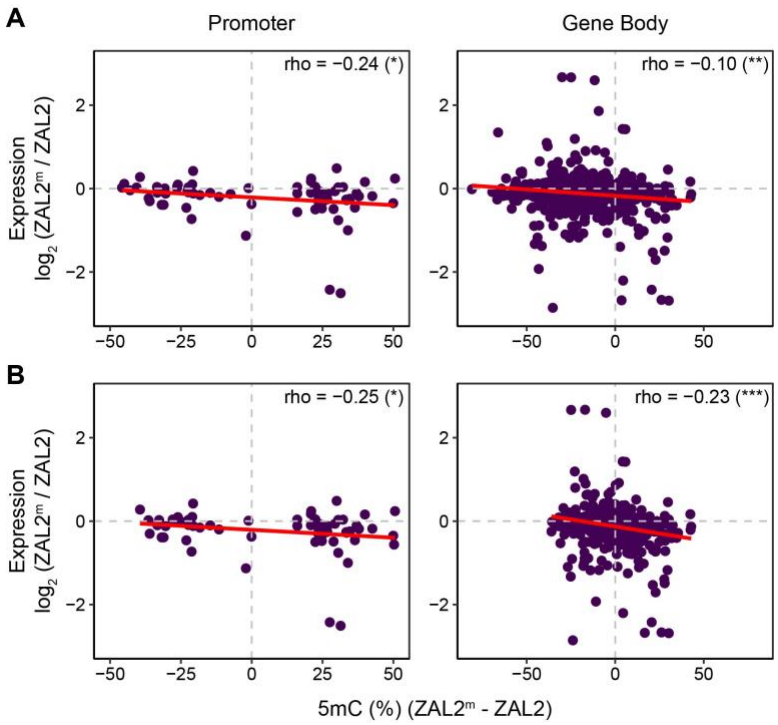

Figure 4.4 Negative relationships between allelic differences in DNA methylation and allelic differences in gene expression for (A) all DMCs and (B) only ZAL2^{m} < ZAL2 and ZAL2^{m} > ZAL2 DMCs. Only promoters and gene bodies that contain more than three DMCs were shown. Allelic differences in DNA methylation across DMCs in a region were averaged. The strength and direction of association were measured by Spearman's rank correlation coefficient, and the relationship was fit with a linear regression line (in red). NS: not significant; *:$P < 0.05$; **:$P < 0.01$; ***:$P < 0.001$.

4.3 Discussion

In this study, we provide evidence that differences in both developmental stages and plumage morphs are associated with the individual variation of the genome-wide DNA methylation patterns in the brains of white-throated sparrows. The comparison between two developmental time points (day-7 chicks and adults) revealed significant age differences in DNA methylation that are widespread across the genome except for the Z chromosome (Figure 4.1). As previous studies of the relationship between DNA

methylation and development/aging usually ignored sex chromosomes, we have yet to understand why age-DMCs are underrepresented on the Z chromosome. Interestingly, age-DMCs are predominantly hypermethylation in adults (Figure 4.2A), which is consistent with the significantly higher expression of DNMTs in adults (Figure 4.2B).

What are the functions of age-related changes in DNA methylation? GO analysis revealed that genes containing age-DMCs in their promoters are enriched for GO terms such as development and cell differentiation (Figure 4.2C). Previous studies have established that DNA hypomethylation through the knockout of DNMTs could cause neural progenitor cells to differentiate into glial cells instead of neurons (Fan et al. 2005; Wu et al. 2010; Murao et al. 2016). Hence, global hypermethylation through DNMT overexpression in adult brains could have functional importance in neurodevelopment in the white-throated sparrow. In support of this idea, a genome-scale DNA methylation study of murine hypothalamus from two post-natal time points found substantial global hypermethylation with age in neuronal but not non-neuronal cells (Li et al. 2014a).

In comparison, differences in DNA methylation between the morphs are almost exclusive to ZAL2/ZAL2^{m} chromosomes. The strong chromosomal enrichment could indicate that genetic differentiation between the two chromosomes might differentially influence methylation of nearby CpGs. Interestingly, the distribution of allelic differences in DNA methylation clearly revealed three classes of DMCs (Figure 4.3A), among which ZAL2^{m} << ZAL2 DMCs have distinct features compared to other allele-DMCs with smaller effect sizes. First, cross-species comparisons revealed that the differences observed in these CpGs are potentially due to ZAL2^{m} hypomethylation rather than ZAL2 hypermethylation (Figure 4.3B). Second, they are less likely than other DMCs to cluster into regions (**Figure C. 2**). Third, their genomic distribution is biased towards TEs and

intronic regions (Figure 4.3C), and accordingly, they are underrepresented in ATAC-seq peaks in which active regulatory regions are enriched (Figure 4.3D).

One hypothesis for the unique genomic distribution of ZAL2^m << ZAL2 DMCs is that they could contribute to the degeneration of ZAL2^m by increasing its chromosomal instability. For example, global hypomethylation in tumors is associated with increased chromosomal instability by facilitating chromosomal rearrangements and activating transposable elements (Rodríguez-Paredes and Esteller 2011). For instance, hypomethylation of the IAP element, one of the most active retrotransposons in the mouse genome, is linked to the higher expression and insertion events of IAP in mouse tumor models (Gaudet et al. 2003; Howard et al. 2007). Furthermore, hypomethylation of repetitive elements such as LINE-1 is associated with a variety of human cancers (Barchitta et al. 2014; Li et al. 2014b; Zelic et al. 2016). However, more data are necessary to examine a direct link between TE methylation and their insertion activity in the ZAL2^m chromosome.

Several lines of evidence support the regulatory roles of the other two classes of allele-DMCs. First, ZAL2^m < ZAL2 DMCs were depleted from ZAL2 ATAC-seq peaks but not in ZAL2^m peaks, and ZAL2^m > ZAL2 DMCs were enriched in ZAL2 peaks but depleted from ZAL2^m peaks (Figure 4.3D). This finding is consistent with an inverse relationship between DNA methylation and transcription activity (presumably strong in ATAC-seq peaks). Second, when the RNA-seq data from the same samples were incorporated, allelic differences in DNA methylation and allelic differences in gene expression are negatively correlated at both promoters and gene bodies (Figure 4.4).

To conclude, our comprehensive epigenetic study in the white-throated sparrow reveals significant effects of developmental time points and plumage morph on DNA

methylation landscapes. We show that age differences in methylation are pervasive and are likely involved in regulating developmental genes, and morph differences are mostly enriched on ZAL2/ZAL2^{m}. Allele-specific methylation analysis unveiled three classes of DMCs, among which two classes may have roles in differentiating gene expression between the two chromosomes. Future studies examining the profiles of histone modifications will allow us to achieve a complete picture of epigenetic regulation in the ZAL2/ZAL2^{m} system.

4.4 Methods

4.4.1 Sample collection

For WGBS and RNA-seq experiments, we collected 12 birds (seven adults [> 1 year old] and five chicks [past hatch day 7]) for our analysis. The hypothalamus was microdissected as previously described (Zinzow-Kramer et al. 2015). For WGS experiments, the livers of the same individuals were collected. For the ATAC-seq experiment, the hypothalamus was microdissected from a white male bird (adult).

4.4.2 WGS library preparation, sequencing, data pre-processing, SNP calling, and identification of fixed differences

Whole genome sequencing libraries were made from DNA extracted from sparrow livers using a QIAGEN DNeasy Blood and Tissue DNA kit. For each sample, 500ng-1µg of DNA was extracted and sheared on a Covaris ultrasonicator to 200-600bp at the Emory Integrated Genomics Core. The DNA fragment ends were repaired, and A-overhangs were added before Nextera barcode adaptors were ligated to the DNA fragments overnight. Finally, the libraries were PCR amplified to increase concentration and enrich

for adaptor-ligated DNA fragments. WGS libraries were sequenced using Illumina HiSeq X Ten with 150 x 2 paired-end reads at Macrogen Clinical Laboratory.

To identify SNPs occupying CpG sites in the reference genome within the sequenced birds, we first trimmed WGS reads of adaptor sequences and of low-quality bases with the parameters "-q 30 -O 1 -m 50 --trim-n --pair-filter any" using cutadapt 1.18 (Martin 2011). Trimmed reads were then aligned to the tan reference genome using Bowtie2 v2.3.4.2 (Langmead and Salzberg 2012) with the --very-sensitive-local option, and the alignment rate was ~95% per sample. Technical duplicates were then discarded by Picard Tools 2.19.0 (https://broadinstitute.github.io/picard/). SNP calling was conducted on clean and aligned reads using GATK 4.0 (McKenna et al. 2010; DePristo et al. 2011; Van der Auwera et al. 2013). Specifically, SNPs were called using Haplotypecaller with the -ERC GVCF option, and joint genotyping of all samples were performed with the GenotypeGVCF. Finally, SNPs with MAF < 0.05, meanDP < 5 and meanDP > 80 were discarded using VCFtools 0.1.15 (Danecek et al. 2011).

With the final set of SNPs, we identified putatively fixed differences between ZAL2 and ZAL2^{m} using the same procedure as described in Chapter 3 (3.4.3). For further alignment of WGBS, ATAC-seq, and RNA-seq data, to minimize potential mapping bias towards the reference genome (ZAL2/ZAL2) caused by differences between ZAL2 and ZAL2^{m}, we constructed a genome with putatively fixed differences masked by Ns in the reference (*N*-masked genome).

4.4.3 WGBS library preparation, sequencing, data pre-processing, and methylation call

WGBS libraries were prepared using a custom protocol. First, DNA was extracted from the brain (hypothalamus region) or testis of white-throated sparrow samples using a

QIAGEN DNeasy Blood and Tissue DNA kit. For each sample, 100 ng - 1 µg of DNA was pooled with 1-5% lambda phage DNA to test for bisulfite conversion efficiency. The DNA samples were then sheared on a Covaris ultrasonicator to 200-600bp. The DNA fragment ends were repaired, and A-overhangs were added before bisulfite compatible adaptors were ligated to the DNA fragments overnight. Then, the DNA fragments were bisulfite-converted and PCR-amplified to increase concentration and enrich for adaptor-ligated DNA fragments. WGBS libraries were then sequenced using Illumina HiSeq X Ten or HiSeq 2500 at Macrogen Clinical Laboratory. At least ~100 million 150 bp x 2 raw reads were generated per sample (Table C. 1).

WGBS reads were trimmed as described in 4.4.2. The trimmed reads were aligned to the *N*-masked reference genome with parameters "--bowtie2 -X 1000" using Bismark v0.20.0 (Krueger and Andrews 2011). The average mapping efficiency of samples was ~70% for all samples (Table C. 1). Third, duplicated reads and non-bisulfite-converted reads were discarded by *deduplicate_bismark* (parameter: -p) and *filter_non_conversion* (parameter: percentage_cutoff 20), respectively. Last, *bismark_methylation_extractor* was run to extract CpG methylation calls.

To obtain bisulfite conversion rates, raw reads were aligned to the phage lambda genome using Bismark (same parameters). As lambda DNA is not methylated and therefore should be completely bisulfite-converted, the percentage of methylated cytosines of lambda DNA is taken as the non-conversion rate. All samples had bisulfite conversion rates above 99.8% (Table C. 1).

To call allele-specific methylation values, SNPsplit 0.3.4 (Krueger and Andrews 2016) was run with parameters "--bisulfite --paired" using fixed differences between ZAL2 and ZAL2^{m}. Then, *bismark_methylation_extractor* was run for allele-separated reads. For

white birds, consistent with the genotype (ZAL2/ZAL2^{m}), the percentage of reads assigned to each chromosome is ~4 - 4.5% (Table C. 1); for tan birds, the percentage of reads assigned to ZAL2 was ~8 - 9% but to ZAL2^{m} 0 - 0.01% (Table C. 1), which was consistent with the genotype (ZAL2 / ZAL2).

After this procedure, the median sequencing depths was at least 9 reads per sample and 4 per allele (Table C. 1). Only CpG sites with at least five reads aligned were retained for further analysis. Finally, as cytosine polymorphisms could hamper accurate methylation call, we excluded any CpGs in the reference genome that were polymorphic within the sequenced samples.

4.4.4 ATAC-seq library preparation, sequencing, data pre-processing, and peak calling

For one sample (hypothalamus of a white male), 10,000 - 200,000 cells from its hypothalamus punch were homogenized in EMEM (Eagle's Minimum Essential Medium) and phosphate-buffered saline. The cells were pelleted in a centrifuge and re-suspended in a lysis buffer made of non-ionic detergent (made in house from Tris, NaCl, $MgCl_2$, IGEPAL CA-630). After cell lysis, nuclei were isolated by centrifugation and added to a tagmentation reaction mix (Illumina Nextera DNA Library Prep Kit, Cat#: FC-121-1030). During tagmentation, the sequencing adapters were inserted into accessible chromatin regions by Tn5 transposase. Adapter-tagmented fragments were purified (Invitrogen Agencourt AMPure XP beads, Cat#: A63880), bar-coded (Illumina Nextera Index Kit, cat#: FC-121-1011), and amplified (Fisher KAPA HiFi HotStart Kit, Cat#: NC0295239). The ATAC-seq libraries were then sequenced by MiSeq (Reagent Kit v3) with 150 cycles (75 bp paired-end reads) in the Molecular Evolution Core at Georgia Tech.

We aligned the trimmed ATAC-seq reads (trimming was performed as described in 4.4.2) to the *N*-masked reference genome using Bowtie2 v2.3.4.2 (parameters: -X 2000 --no-mixed --no-discordant) (Langmead and Salzberg 2012), which allowed a maximal insert size of 2 Kb between paired reads and discarded any unmapped or discordant alignment. The mapping efficiency for this sample was 82.13%. The aligned reads were then deduplicated using markdup of samtools 1.7 (Li et al. 2009). As a result, we obtained 23 million clean mapped reads. To assign reads to ZAL2 and ZAL2^{m}, SNPsplit 0.3.4 (Krueger and Andrews 2016) was run with parameters "--paired" using fixed differences between ZAL2 and ZAL2^{m}. Further, ZAL2 and ZAL2^{m} ATAC-seq peaks were called using MACS2 version 2.1.1.20160309 (Zhang et al. 2008) with '-g 1.1e+9 -f BAMPE -p 0.01 -B --SPMR --nomodel' options.

4.4.5 RNA-seq library preparation, sequencing, data processing, and differential expression analysis

RNA extraction and library preparation of the female samples was performed as previously described (Zinzow-Kramer et al. 2015). The libraries were then sequenced on the HiSeq 4000 at 150 PE reads to ~40 million reads per sample. RNA-seq raw reads were trimmed as described in 4.4.2 and then aligned to the *N*-masked genome by HISAT2 2.1.0 (Kim et al. 2015). Secondary alignments were filtered by SAMtools 1.7 (Li et al. 2009) to ensure that only primary alignments were retained. SNPsplit 0.3.4 (Krueger and Andrews 2016) was run to assign reads to ZAL2 or ZAL2^{m} for the white samples and to filter out reads without fixed differences in the tan samples. Expression levels (raw read counts) were then quantified by StringTie v1.3.4d (Pertea et al. 2015).

To identify genes that were differentially expressed between ZAL2 and ZAL2^{m}, we normalized libraries with the size factors generated in the morph comparison step and

identified differential expression with ‘design = ~ age + allele’ (age as the adjusted covariate) using the DESeq2 1.22.2 package (Love et al. 2014) in R 3.5 (R Core Team 2019).

4.4.6 Analysis of differential DNA methylation

Differentially methylated CpGs between age groups, sexes, and morphs (or alleles) were detected by DSS 2.30.1 (Wu et al. 2015) under the default setting, with one variable as the independent variable and the other two as adjusted covariates. CpGs with FDR-corrected p-values less than 0.05 and absolute values of differences in methylation greater than 10% were defined as DMCs. Bedtools v2.28.0 (Quinlan and Hall 2010) was run to assign DMCs to different gene features. When DMCs are within multiple gene features, we prioritize the assignment in the following order: upstream (10 Kb upstream of TSS), exons, introns, downstream (10 Kb downstream of TES), TEs and intergenic region. The distribution of all CpGs on ZAL2 and ZAL2^{m} was employed as the control.

4.4.7 Principal component analysis

We stored DNA methylation data generated from all samples as an methylrawDB object using methylkit 1.9.4 (Akalin et al. 2012). The object was then converted into a percent methylation matrix, with only CpG sites with more than five reads in all samples retained. PCA analysis was performed using the PCASamples function in methylkit (parameter: obj.return = T). The returned prcomp result was used to plot sample clusters with the autoplot function in ggfortify 0.4.5 (Tang et al. 2016).

4.4.8 Transposable element annotation

We adopted both *de novo* and homology-based approaches to annotate repetitive sequences in the reference genome. First, *de novo* discovery of TEs were performed by RepeatModeler 1.0.9 (Smit and Hubley 2008-2015). The generated library was merged with the avian Repbase library (20181026 version), which was used to annotate TEs in the reference genome using RepeatMasker 4.0.9 (parameters: -xsmall -s -nolow -norna -nocut) (Smit et al. 2013-2015).

4.4.9 Cross-species whole genome alignment and DNA methylation comparison

To assign scaffolds of the reference white-throated sparrow genome to zebra finch chromosomes or to compare CpG methylation levels between sparrow and the outgroup species (great tit), we aligned the sparrow reference genome to zebra finch (Taeniopygia_guttata-3.2.4) and great tit (Parus_major1.1) reference genomes using minimap2-2.16 (parameters: --secondary=no -c) (Li 2018). Only confident alignments, defined by the highest mapping score (MAPQ=60), were retained. The paftools liftover program of minimap2 was then run to find dinucleotides in the great tit and chicken genome that were orthologous to CpG sites in the sparrow genome. Using the brain methylation data of the great tit and chicken generated in Chapter 2, we obtained fractional methylation levels for shared CpGs in the other two species (sparrow-tit: 436 CpGs; sparrow-chicken: 87 CpGs).

4.5 Acknowledgements

This work is a collaboration between Dr. Soojin Yi's Lab at Georgia Tech and Dr. Donna Maney's Lab at Emory University. We thank members of the Maney Lab for providing sparrow samples and conducting RNA-seq experiments (Dr. Kathleen Grogan).

We thank members of the Yi Lab for performing WGS, WGBS and ATAC-seq experiments (Thomas Layman) as well as WGS data processing and SNP calling (Hyeonsoo Jeong).

CHAPTER 5

CONCLUSIONS

The common path of the evolution of sex chromosomes in most species is a step-wise cessation of recombination between the proto-sex chromosomes (Charlesworth 1978; Lahn and Page 1999). My dissertation is focused on two sex chromosome(-like) systems: the avian Z and W chromosomes that evolved more than 100 million years ago and the young ZAL2 and ZAL2^{m} chromosomes in the white-throated sparrow that emerged only 2-3 million years ago. The two time points offer a powerful opportunity to study the regulatory evolution of sex chromosome(-like) systems in action.

For the old avian Z and W chromosomes, W is largely degenerated, and the vast majority of Z-linked genes do not have W counterparts. Because of incomplete dosage compensation, ZW females generally have lower gene expression than ZZ males. Yet, dosage regulation of some genes may be essential in females, and previous studies have indicated a role of DNA methylation in local dosage compensation on the chicken Z chromosome (Melamed and Arnold 2007). However, no information on the chromosome-wide epigenetic difference in avian sex chromosomes was available until my research. To fill this critical gap of knowledge, in Chapter 2, I investigated chromosome-wide sex differences in DNA methylation on the Z chromosome in the chicken and white-throated sparrow.

Consistent with incomplete dosage compensation, DNA methylation, which is a useful tool in mammalian X inactivation, does not differ globally between males and females in both species. In chicken, I discovered the convergent evolution of two MHM regions (one was previously identified by Teranishi et al. [2001]). At both loci, DNA

methylation effectively blocks transcription in males by drastically reducing chromatin accessibilities, and lncRNAs are expressed almost exclusively in females. Genes in the vicinity of MHMs display notably reduced male-to-female expression ratios, which is likely due to the binding of lncRNAs in females. This study demonstrates that, at least twice in their evolutionary history, chicken (and potentially other Galloanserae species because they also contain MHM sequences) employed DNA methylation as a mechanism to solve dosage problems for genes potentially useful to females. In the white-throated sparrow, however, I did not detect any MHM regions. This could indicate that dosage regulation in the latter may be achieved on a gene-by-gene basis, rather than a regional basis.

One question that remains unanswered is how sex differences in the DNA methylation of MHMs are achieved. I hypothesize that some “switches” on the W chromosome may control the methylation status of MHMs. For example, Teranishi et al. (2001) compared DNA methylation levels of MHM1 between triploids and standard males/females. They found that the methylation levels of MHM1 were almost equally low in ZZW and standard ZW females, and DNA methylation levels were almost equally high in ZZZ and standard ZZ males. These observations suggest that the presence/absence of a W chromosome, instead of the number of Z chromosomes, determines the methylation status of MHM1 in chicken. For future work, we can conduct targeted bisulfite sequencing of both MHMs in triploids and standard males/females to validate this finding and extend it to MHM2. Identifying these potential switches (if exist) can be challenging. However, comparing the methylation levels of MHMs before and after knocking out W-linked genes could be a good starting point.

In contrast to the avian sex chromosomes in which W chromosome is mostly decayed, ZAL2 and ZAL2^{m} in the white-throated sparrow are still similar in sizes. Rare but otherwise healthy ZAL2^{m}/ZAL2^{m} homozygotes have been observed in nature, indicating

that the ZAL2^m is at the very early stage of degeneration, if at all. In Chapter 3, I examined the degree of genetic and expression divergence between the ZAL2 and ZAL2^m chromosomes as well as the degree of degeneration of ZAL2^m. The ZAL2 chromosome (present in birds of both tan [ZAL2/ZAL2] and white [ZAL2/ZAL2^m] morphs) and the non-recombining ZAL2^m chromosome (present only in birds of the white [ZAL2/ZAL2^m] morph) are distinguished by nucleotide divergence of ~1%. The ZAL2^m chromosome exhibits signatures of degeneration, including slight accumulation of deleterious mutations at functional sites and subtle pseudogenization.

Surprisingly, despite the weak genetic differentiation and degeneration, more than 40% of genes exhibit differential expression in the brain. Among these genes, significantly more genes are ZAL2-biased, which is consistent with ZAL2^m degeneration. Moreover, patterns of gene expression are consistent with dosage compensation re-balancing expression levels between the two morphs. Specifically, ZAL2-linked genes appear to increase their expression in white birds to counter the expression loss of their ZAL2^m counterparts. These results indicate that the regulatory evolution can precede large-scale genetic differences for this incipient system.

We have yet to understand how dosage compensation is achieved in the ZAL2/ZAL2^m system. In both fruit flies and green anoles, acetylation of H4K16 relaxes the chromatin and doubles the expression of X-linked genes in males (Conrad and Akhtar 2011; Marin et al. 2017). Interestingly, MHM1 also exhibited a strong enrichment of H4K16ac in female chickens (Bisoni et al. 2005); hence, acetylation of H4K16 may serve as a common mechanism for dosage compensation in many sex chromosome systems. In future studies, by performing ChIP-seq experiments, we can test whether this histone mark has a similar role in the dosage compensation in the ZAL2/ZAL2^m system.

One possible mechanism that could have generated the observed expression divergence between ZAL2 and ZAL2^{m} is their epigenetic differentiation. In Chapter 4, I examined DNA methylation profiles of tan and white birds from two developmental time points. Although age effects on DNA methylation are pervasive across the genome, morph effects are almost exclusive to the ZAL2/2^{m} chromosomes. The analysis of allelic methylation revealed three classes of DMCs. By incorporating ATAC-seq and RNA-seq data, I found that DMCs with moderate differences in DNA methylation may partially explain the expression differentiation between ZAL2 and ZAL2^{m}. Even though we have yet to understand the functional consequence of extremely hypomethylated CpGs on ZAL2^{m}, this class of DMCs has unique genomic distribution: they are depleted from active regulatory regions and are enriched in TEs. One hypothesis is that they may contribute to the degeneration of ZAL2^{m} by activating transposable elements. Future studies can test whether the TEs containing these DMCs have more copies on ZAL2^{m} than ZAL2 using experimental and computational approaches.

My comprehensive genetic, transcriptomic, and epigenetic studies help to expand our knowledge about the regulatory evolution of both old and nascent sex chromosome(-like) systems in birds. Although the two systems differ drastically in their extents of genetic divergence, both display intriguing expression patterns potentially achieved by epigenetic regulation. I hope my research stimulates interest in epigenetic studies of non-mammalian species in the scientific community. With rapid advances in epigenetic sequencing technologies, future studies will help us to gain a deeper understanding of how different genetic and environmental factors can shape epigenetic landscapes as well as how DNA methylation and other epigenetic mechanisms interweave to mediate dosage compensation and phenotypic dimorphism in birds.

APPENDIX A

SUPPLEMENTARY MATERIAL FOR CHAPTER 2

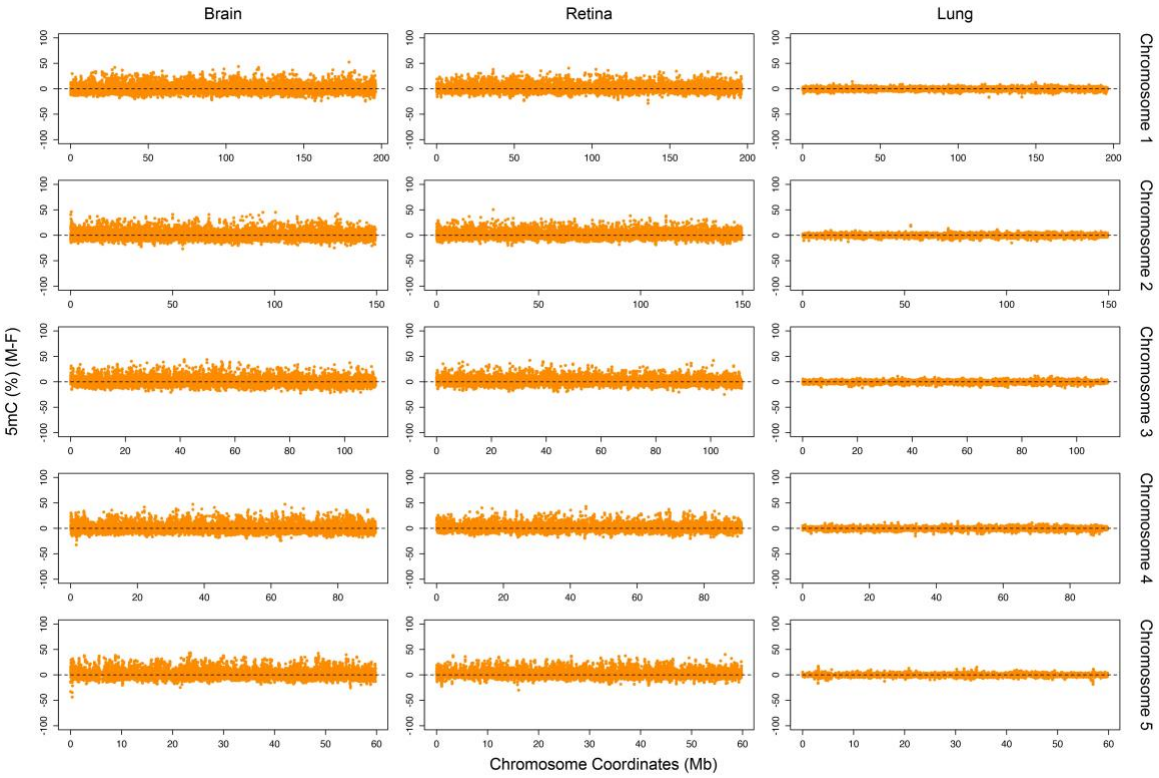

Figure A. 1 The absence of notable outliers of sex differences in methylation on autosomes. Only the first five autosomes (macrochromosomes) of the chicken genome are shown, but outliers are also absent from other autosomes (intermediate chromosomes and microchromosomes). Methylation values were plotted using a 10 Kb window size with a 1 Kb step size.

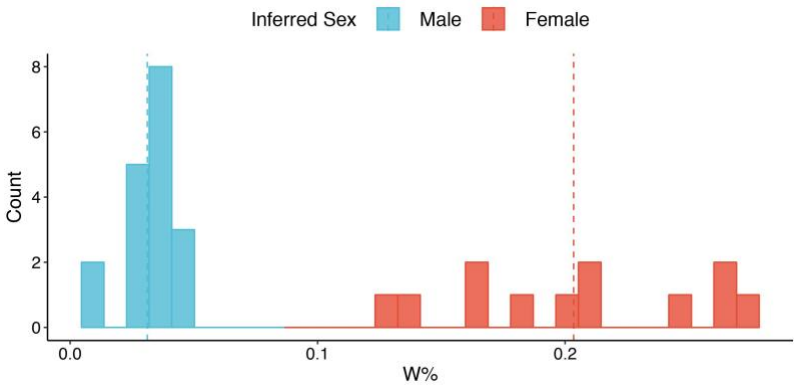

Figure A. 2 The distribution of percent W reads (W%) for several WGBS and ATAC-seq samples without sex information. Percent W reads was calculated as the percentage of reads mapped to the W chromosome out of all mapped reads for each sample. The distribution reveals two groups of samples: samples from the group with high W% were defined as females and samples from the other group were defined as males. Dashed lines represent the mean values.

Table A. 1 Sample information for RNA-seq, ATAC-seq, and WGBS data used in this study.

Data Type	Species	Tissue/Cell type	Age	Accession #	# of samples & sex	Read type	Source
RNA-seq	Chicken	Blastoderm	12hr	PRJNA171809	2M; 2F	Paired-end	(Ayers et al. 2013)
	Chicken	Gonad	E4.5	PRJNA171809	2M; 2F	Paired-end	(Ayers et al. 2013)
	Chicken	Gonad	E6	PRJNA171809	2M; 2F	Paired-end	(Ayers et al. 2015)
	Chicken	Brain	E18	PRJEB8390	5M; 5F	Paired-end	(Uebbing et al. 2015)
	Chicken	Bursa	E18	PRJEB8390	5M; 5F	Paired-end	(Uebbing et al. 2015)
	Chicken	Heart	E18	PRJEB8390	5M; 5F	Paired-end	(Uebbing et al. 2015)
	Chicken	Kidney	E18	PRJEB8390	5M; 5F	Paired-end	(Uebbing et al. 2015)
	Chicken	Liver	E18	PRJEB8390	4M; 4F	Paired-end	(Uebbing et al. 2015)
	Chicken	Lung	E18	PRJEB8390	5M; 5F	Paired-end	(Uebbing et al. 2015)
	Chicken	Muscle	E18	PRJEB8390	5M; 5F	Paired-end	(Uebbing et al. 2015)
	Chicken	Spleen	E18	PRJEB8390	5M; 4F	Paired-end	(Uebbing et al. 2015)
	Chicken	Gonad	E18	PRJEB8390	5M; 5F	Paired-end	(Uebbing et al. 2015)
	Chicken	Liver	E19	PRJNA284655	4M; 4F	Paired-end	(Zimmer et al. 2016)
	Chicken	Heart	E19	PRJNA284655	4M; 4F	Paired-end	(Zimmer et al. 2016)
	Chicken	Spleen	E19	PRJNA284655	4M; 4F	Paired-end	(Zimmer et al. 2016)
	Chicken	Gonad	E19	PRJNA284655	4M; 4F	Paired-end	(Zimmer et al. 2016)
	Chicken	Brain	Adult	PRJNA143627	1M; 1F	Single-end	(Brawand et al. 2011; Julien et al. 2012)
	Chicken	Heart	Adult	PRJNA143627	1M; 1F	Single-end	(Brawand et al. 2011; Julien et al. 2012)
	Chicken	Kidney	Adult	PRJNA143627	1M; 1F	Single-end	(Brawand et al. 2011; Julien et al. 2012)
	Chicken	Liver	Adult	PRJNA143627	1M; 1F	Single-end	(Brawand et al. 2011; Julien et al. 2012)
	Chicken	Brain	Adult	PRJNA381064	2M; 1F	Single-end	(Marin et al. 2017)
	Chicken	Heart	Adult	PRJNA381064	2M; 2F	Single-end	(Marin et al. 2017)
	Chicken	Kidney	Adult	PRJNA381064	2M; 2F	Single-end	(Marin et al. 2017)
	Chicken	Liver	Adult	PRJNA381064	2M; 2F	Single-end	(Marin et al. 2017)
	Chicken	Gonad	Adult	PRJNA381064	2M; 2F	Single-end	(Marin et al. 2017)
	Chicken	Brain	Adult	PRJEB4677	1M; 1F	Paired-end	(Chickspress 2015)
	Chicken	Gonad	Adult	PRJEB4677	1M; 1F	Paired-end	(Chickspress 2015)
	Chicken	Gonad	Adult	PRJNA186646	1M; 2F	Paired-end	(Necsulea et al. 2014)
	Blue tit	Brain	Adult	PRJNA284903	5M; 5F	Paired-end	(Mueller et al. 2015)
	Ostrich	Brain	Adult	SRP012236	3M; 3F	Paired-end	(Adolfsson and Ellegren 2013)
WGBS	Chicken	Brain	E18	PRJNA389197	1M; 2F	Paired-end	(Lee et al. 2017)
	Chicken	Retina	E18	PRJNA389197	1M; 2F	Paired-end	(Lee et al. 2017)
	Chicken	Lung	3wk	PRJNA245100	2M; 2F	Paired-end	(Li et al. 2015)
	Chicken	Breast muscle	20wk	PRJNA352686	1F	Paired-end	(Zhang et al. 2017)
	Chicken	Breast muscle	Adult	PRJNA352686	1F	Paired-end	(Zhang et al. 2017)
	Great tit	Brain	Adult	PRJNA208335	1M	Paired-end	(Derks et al. 2016; Laine et al. 2016)
	Sparrow	Brain	Adult	PRJNA540850	1M; 1F	Paired-end	This study
	Human	Brain	Adult	PRJNA158561	2M; 1F	Single-end	(Zeng et al. 2012)
ATAC-seq	Chicken	Forelimb	E4.5	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Hindlimb	E4.5	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Flight Muscle	E9	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Inferior Sternum	E9	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Superior Sternum	E9	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Flight Muscle	E10	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Keel	E10	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Full Sternum	E10	PRJNA433154	2M; 1F	Paired-end	(Sackton et al. 2018)
	Chicken	Liver	Adult	PRJEB27111	2M; 2F	Paired-end	(Foissac et al. 2018)
	Chicken	$CD4^+$ T cells	Adult	PRJEB27111	2M; 2F	Paired-end	(Foissac et al. 2018)

Table A. 2 The coordinates of 26 Z and W gametologs used in this study.

Gene Symbol	W Accession	W Start	W End	W Strand	Z Accession	Z Start	Z End	Z Strand
ATP5A1	NC_006126.5	1391533	1453513	+	NC_006127.4	2151416	2159390	+
C18orf25	NC_006126.5	1460566	1669620	+	NC_006127.4	2087422	2131172	-
CHD1	NC_006126.5	4989932	5105612	-	NC_006127.4	51274317	51322737	+
GOLPH3	NC_006126.5	3970457	4106230	+	NC_006127.4	9664934	9697216	-
GPBP1	NC_006126.5	4494841	4570559	+	NC_006127.4	17517405	17555385	+
HINT1	NC_006126.5	1839400	1896965	+	NC_006127.4	44899756	44903752	+
HNRNPK	NC_006126.5	4703885	4724023	-	NC_006127.4	40076888	40095691	-
KCMF1	NC_006126.5	6346825	6447458	+	NC_006127.4	53774220	53829526	-
KIAA0427/CTIF	NC_006126.5	2667781	2705020	-	NC_006127.4	1278266	1374473	-
MIER3	NC_006126.5	4214767	4365372	-	NC_006127.4	17476327	17499064	-
NEDD4L	NC_006126.5	3023271	3421031	-	NC_006127.4	613150	743160	+
NIPBL	NC_006126.5	4788268	4948399	+	NC_006127.4	11295476	11454770	+
RASA1	NC_006126.5	166182	227499	+	NC_006127.4	61010336	61071982	-
RPL17	NC_006126.5	2397226	2407148	+	NC_006127.4	1078253	1081593	+
SMAD2	NC_006126.5	6650804	6715919	+	NC_006127.4	1515826	1555787	+
SMAD7	NC_006126.5	2602132	2628699	+	NC_006127.4	1241970	1271132	+
SPIN1	NC_006126.5	5939316	6052602	+	NC_006127.4	43172671	43247852	+
ST8SIA3	NC_006126.5	6538043	6552493	+	NC_006127.4	435916	439939	-
SUB1/PC4/RPTC15L	NC_006126.5	5302776	5319796	+	NC_006127.4	9801318	9816323	+
TXNL1	NC_006126.5	3494254	3517698	+	NC_006127.4	593165	607799	+
UBAP2	NC_006126.5	2474	48083	+	NC_006127.4	7358767	7496899	-
UBE2R2	NC_006126.5	67936	141538	-	NC_006127.4	7304115	7355704	+
VCP	NC_006126.5	3607608	3662599	-	NC_006127.4	8540992	8562186	-
ZFR/ZFR2	NC_006126.5	5153701	5259636	-	NC_006127.4	9752237	9793319	-
ZNF532	NC_006126.5	2138679	2216237	+	NC_006127.4	852624	892107	+
ZSWIM6	NC_006126.5	366326	817288	-	NC_006127.4	19094888	19206550	+

APPENDIX B

SUPPLEMENTARY MATERIAL FOR CHAPTER 3

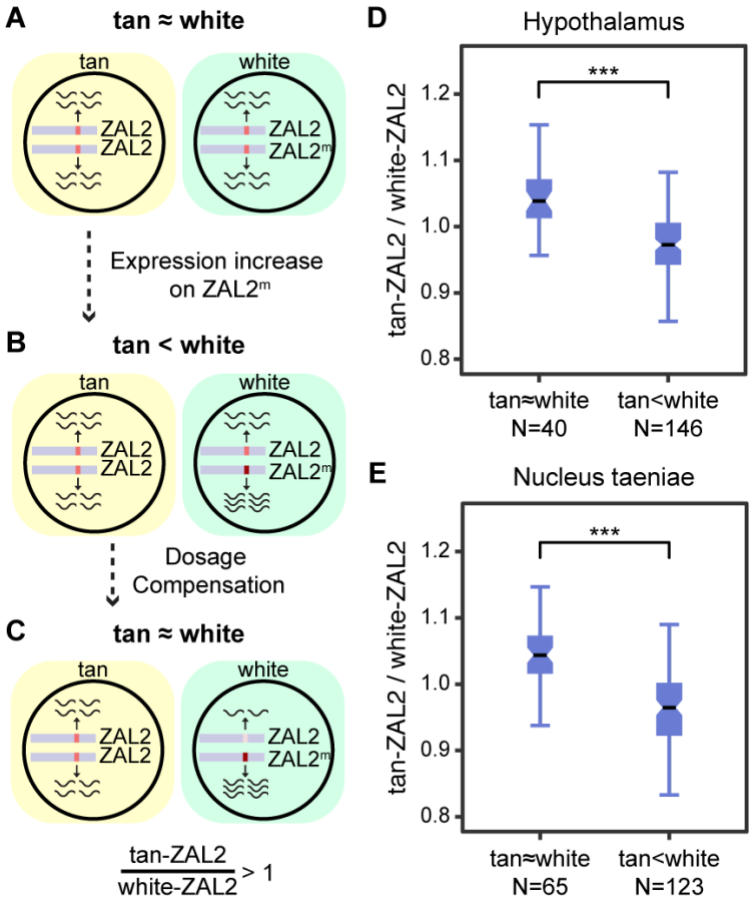

Figure B. 1 Potential dosage compensation for ZAL2^m-biased genes. (A) Before degeneration, expression dosage (black waves) is similar between ZAL2 and ZAL2^m and between tan and white. **(B)** If the expression of ZAL2^m alleles is increased and there is no dosage compensation, white individuals should show overexpression. **(C)** The dosage between the morphs may be re-balanced via downregulation of the ZAL2 allele in white birds. Consequently, the expression of the ZAL2 allele should be lower in white than tan birds (tan-ZAL2 / white-ZAL2 > 1). **(D-E)** Levels of compensation (measured by tan-ZAL2 / white-ZAL2) are significantly elevated for tan ≈ white (dosage-compensated) genes compared with non-dosage compensated genes (tan < white).

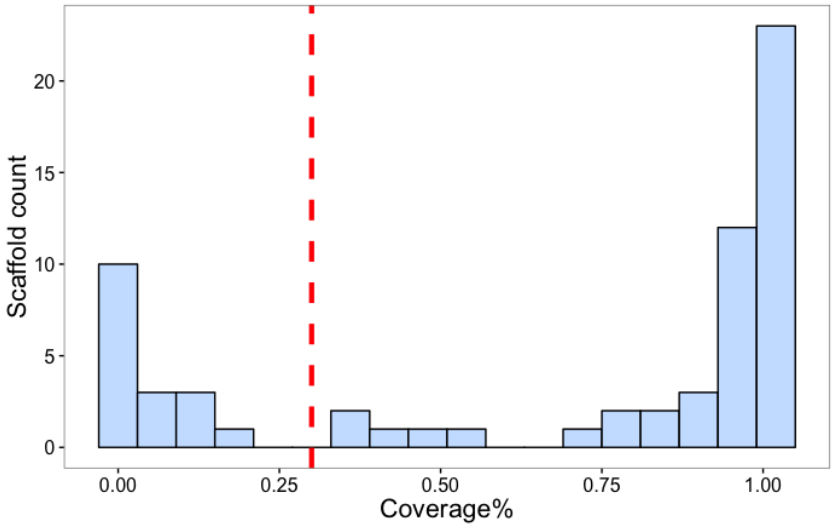

Figure B. 2 The distribution of percent coverage (Coverage %) for scaffolds in the tan reference genome. Coverage % was calculated as the length of the region that could be mapped to the TGU3 chromosome, divided by the total length of that scaffold. The cutoff for coverage % by our criteria is shown as the red dashed line. Only scaffolds >10 Kb are included.

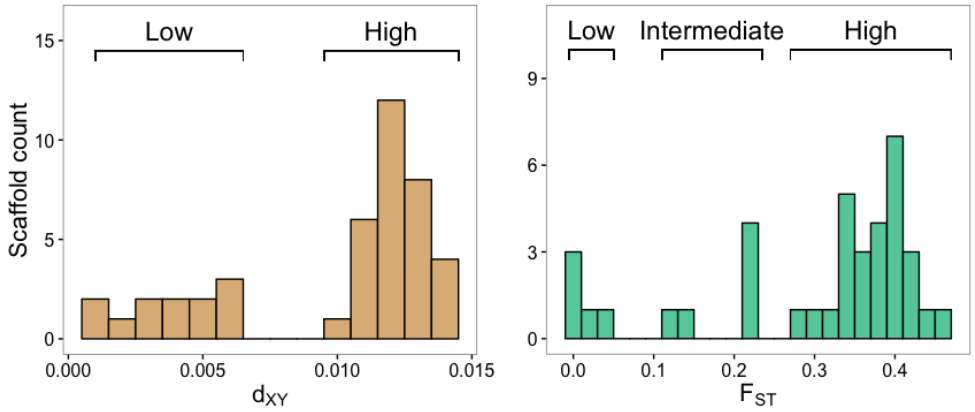

Figure B. 3 Bimodal patterns of divergence for pairwise nucleotide divergence and degrees of population differentiation. The graphs show the distribution of average pairwise nucleotide divergence (d_{XY}) between ZAL2 and ZAL2^{m} chromosomes (left panel) and degrees of population differentiation (F_{ST}) between white and tan samples over ZAL2 scaffolds (right panel). Cutoffs to distinguish between 'Low' and 'High' (as well as 'Intermediate' in the case of F_{ST}) of the two measures were determined by the clear divisions in the distributions. Thus, low d_{XY} is in the range of [0.00080, 0.00588], and high d_{XY} corresponds to [0.00968, 0.01431]. Similarly, low F_{ST} is in the range of [0.00448, 0.01420], intermediate F_{ST} is [0.11821, 0.22853], and high F_{ST} is [0.28586, 0.45355].

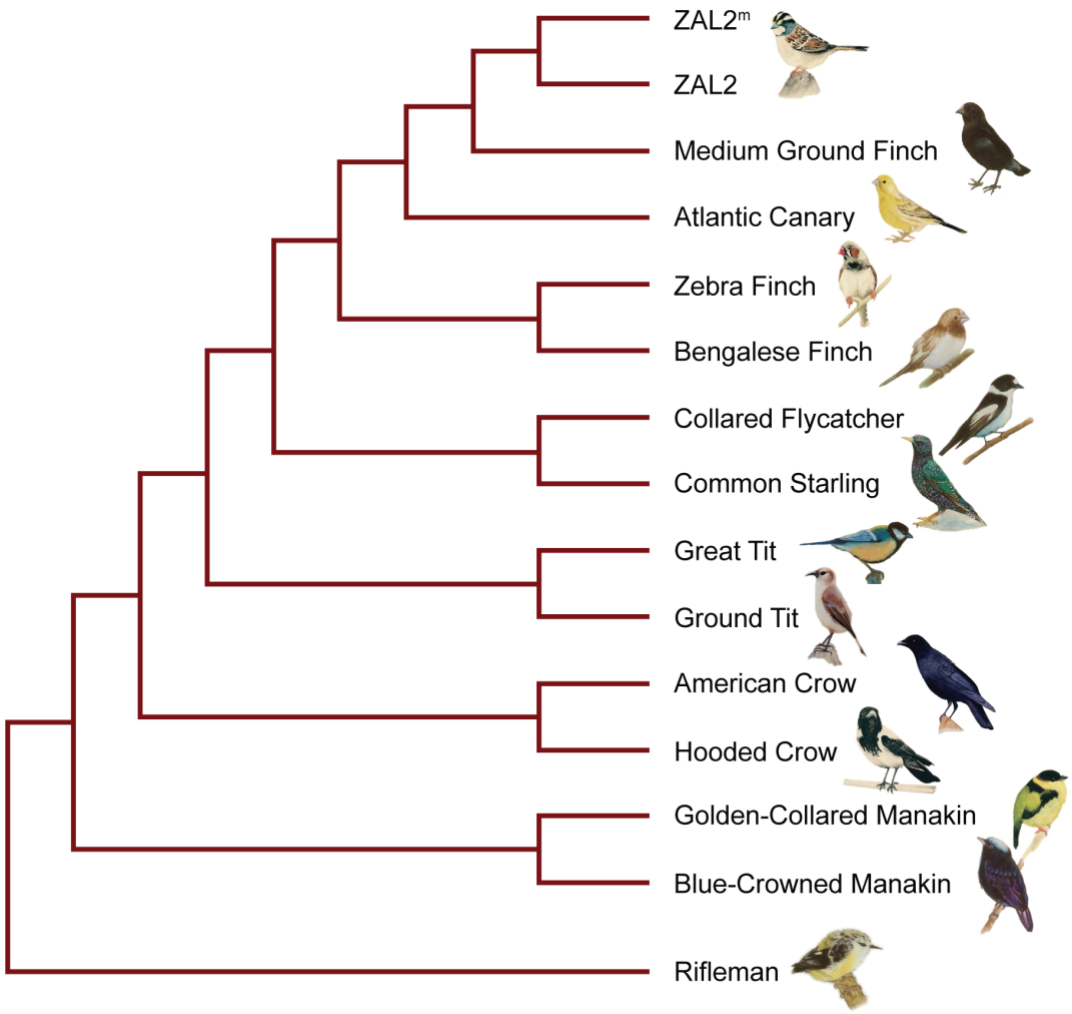

Figure B. 4 The phylogeny of the passeriform species used in our analysis. This species tree was inferred from several avian phylogeny studies (Jetz et al. 2012; Jarvis et al. 2014; Jetz et al. 2014; Prum et al. 2015).

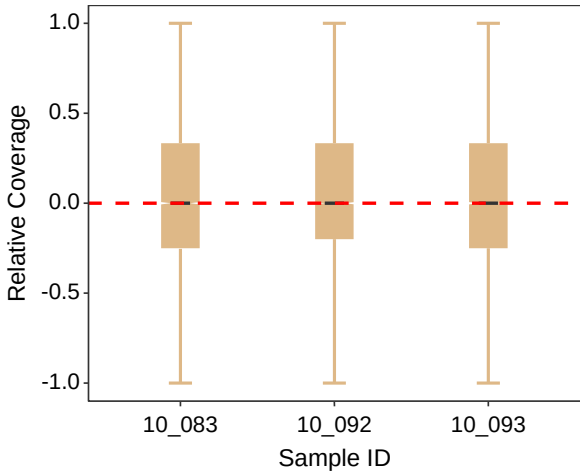

Figure B. 5 Evidence against mapping bias. Our SNP *N*-masking approach prevented mapping bias toward the reference ZAL2 alleles. The Y-axis represents normalized differences of per base coverage between ZAL2 and ZAL2^{m} for each fixed difference (specifically, (ZAL2 coverage − ZAL2^{m} coverage) / (ZAL2 coverage + ZAL2^{m} coverage)). No significant bias towards the ZAL2 allele was detected across all three samples (paired t-test on $\log_2$-transformed counts, $P > 0.05$).

APPENDIX C

SUPPLEMENTARY MATERIAL FOR CHAPTER 4

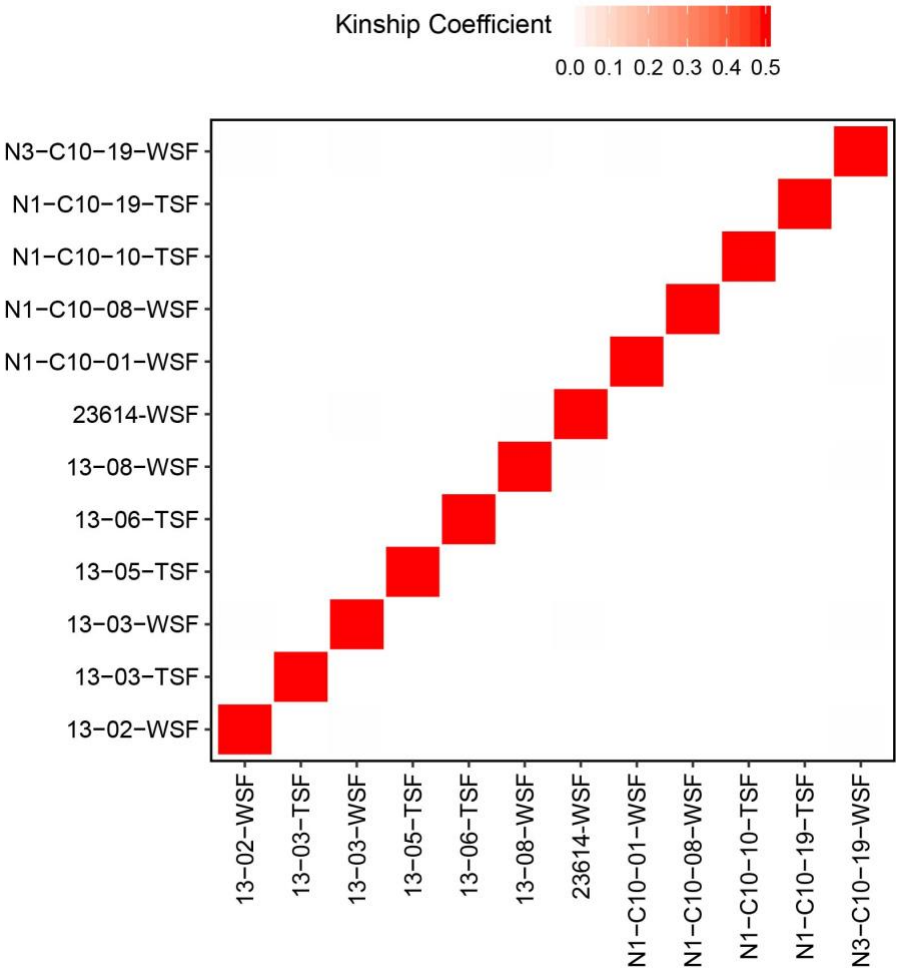

Figure C. 1 Pairwise kinship coefficients of WGBS samples used in this study. Kinship coefficients were calculated using KING. Negative values were set to zeros. The coefficients range from 0 (unrelated) to 0.5 (duplicates/monozygotic twins).

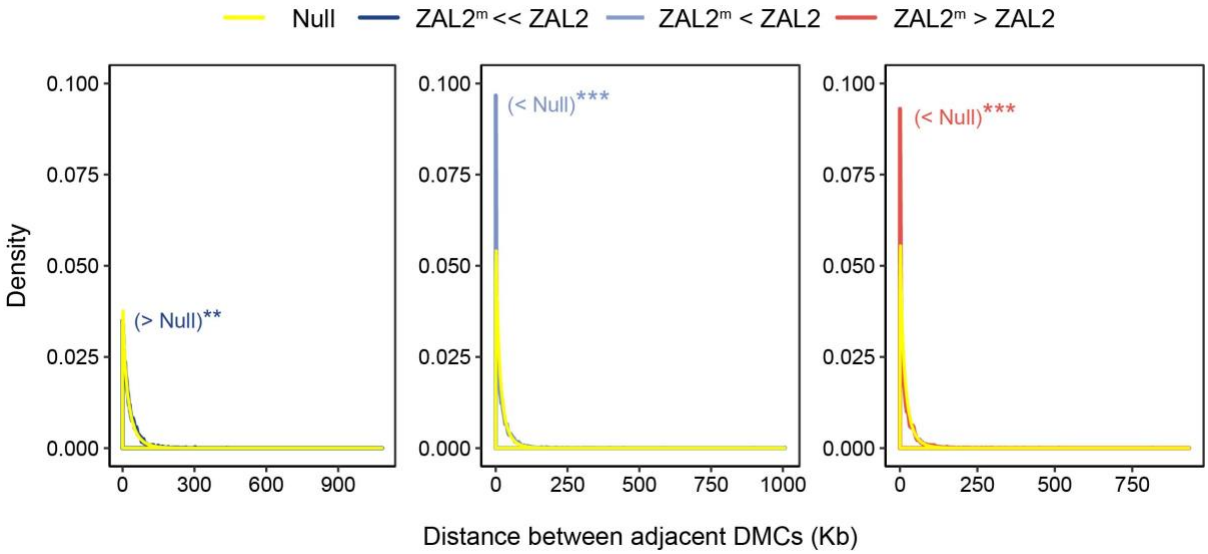

Figure C. 2 Distribution of distance between adjacent DMCs. To obtain null distributions, we randomly chose the same number of CpGs per group from the chromosome and recorded distances between adjacent CpGs. This step was repeated 100 times. Paired Mann–Whitney *U* test was used to compare the null distributions with real distributions ('> Null': distances between adjacent CpGs are greater than expected; '< Null': distances between adjacent CpGs are smaller than expected; **:$P < 0.01$, ***:$P < 0.001$).

Table C. 1 WGBS sample pre-processing stats.

Sample ID	Age	Morph	Sex	Tissue	Raw reads	Mapping efficiency (%)	Library duplication rate (%)	Bisulfite conversion rate (%)	Total clean aligned reads	ZAL2 reads	ZAL2 reads (%)	ZAL2^m reads	ZAL2^m reads (%)	Median depth (Total)	Median depth (2)	Median depth (2^m)
N1-C10-01	chick	white	F	Hyp	431,301,497	69.2	33.98	99.8	147,714,025	6,539,090	4.43	5,975,572	4.05	29	12	11
N1-C10-08	chick	white	F	Hyp	443,183,601	68.2	50.81	99.8	129,239,725	5,648,773	4.38	5,173,722	4.01	26	10	9
N3-C10-19	chick	white	F	Hyp	453,064,974	66	59.28	99.8	105,757,536	4,628,887	4.38	4,229,833	4.01	21	8	7
23614-WSF	adult	white	F	Hyp	456,487,223	72.6	17.48	99.8	241,822,969	10,908,377	4.56	9,915,458	4.14	48	20	18
WSF-13-02	adult	white	F	Hyp	413,525,337	70.3	26.92	99.8	186,511,833	8,210,248	4.45	7,458,448	4.04	38	16	14
WSF-13-03	adult	white	F	Hyp	380,468,028	70.3	51.4	99.8	114,186,096	5,222,263	4.62	4,344,904	3.84	22	9	7
WSF-13-08	adult	white	F	Hyp	438,146,751	71.4	16.27	99.8	231,819,927	10,304,477	4.48	9,379,039	4.08	50	21	18
N1-C10-10	chick	tan	F	Hyp	143,904,871	72.5	17.03	99.8	72,424,331	6,336,970	8.77	3,406	0	13	NA	NA
N1-C10-19	chick	tan	F	Hyp	127,913,887	74.8	10.69	99.8	74,391,203	6,553,425	8.83	3,217	0	13	NA	NA
TSF-13-03	adult	tan	F	Hyp	124,791,096	75.4	15.87	99.8	67,260,923	6,045,486	9.10	2,727	0	11	NA	NA
TSF-13-05	adult	tan	F	Hyp	131,106,615	73.9	17.32	99.8	66,862,821	5,962,661	9.00	2,748	0	11	NA	NA
TSF-13-06	adult	tan	F	Hyp	142,979,273	73	13.45	99.8	77,578,167	6,863,103	8.94	3,265	0	15	NA	NA

REFERENCES

Adolfsson S, Ellegren H. 2013. Lack of dosage compensation accompanies the arrested stage of sex chromosome evolution in ostriches. *Molecular Biology and Evolution* **30**: 806-810.

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE. 2012. methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. *Genome Biol* **13**: R87.

Anders S, Pyl PT, Huber W. 2015. HTSeq — a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**: 166-169.

Asnicar MA, Smith DP, Yang DD, Heiman ML, Fox N, Chen Y-F, Hsiung HM, Köster A. 2001. Absence of cocaine- and amphetamine-regulated transcript results in obesity in mice fed a high caloric diet. *Endocrinology* **142**: 4394-4400.

Ayers KL, Davidson NM, Demiyah D, Roeszler KN, Grützner F, Sinclair AH, Oshlack A, Smith CA. 2013. RNA sequencing reveals sexually dimorphic gene expression before gonadal differentiation in chicken and allows comprehensive annotation of the W-chromosome. *Genome Biol* **14**: R26.

Ayers KL, Lambeth LS, Davidson NM, Sinclair AH, Oshlack A, Smith CA. 2015. Identification of candidate gonadal sex differentiation genes in the chicken embryo using RNA-seq. *BMC Genomics* **16**: 704.

Bachtrog D. 2013. Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nat Rev Genet* **14**: 113-124.

Bachtrog D, Mank JE, Peichel CL, Kirkpatrick M, Otto SP, Ashman T-L, Hahn MW, Kitano J, Mayrose I, Ming R et al. 2014. Sex determination: why so many ways of doing it? *PLoS Biol* **12**: e1001899.

Barchitta M, Quattrocchi A, Maugeri A, Vinciguerra M, Agodi A. 2014. LINE-1 hypomethylation in blood and tissue samples as an epigenetic marker for cancer risk: a systematic review and meta-analysis. *PLoS One* **9**: e109478-e109478.

Bellott DW, Skaletsky H, Cho TJ, Brown L, Locke D, Chen N, Galkina S, Pyntikova T, Koutseva N, Graves T et al. 2017. Avian W and mammalian Y chromosomes convergently retained dosage-sensitive regulators. *Nat Genet* **49**: 387-394.

Bird A. 1992. The essentials of DNA methylation. *Cell* **70**: 5-8.

Bird A. 2002. DNA methylation patterns and epigenetic memory. *Gene Dev* **16**: 6-21.

Bisoni L, Batlle-Morera L, Bird AP, Suzuki M, McQueen HA. 2005. Female-specific hyperacetylation of histone H4 in the chicken Z chromosome. *Chromosome Res* **13**: 205-214.

Brawand D, Soumillon M, Necsulea A, Julien P, Csárdi G, Harrigan P, Weier M, Liechti A, Aximu-Petri A, Kircher M et al. 2011. The evolution of gene expression levels in mammalian organs. *Nature* **478**: 343.

Brockdorff N. 2018. Local tandem repeat expansion in Xist RNA as a model for the functionalisation of ncRNA. *Non-Coding RNA* **4**.

Brockdorff N, Turner BM. 2015. Dosage compensation in mammals. *Cold Spring Harb Perspect Biol* **7**: a019406.

Bujo H, Hermann M, Lindstedt KA, Nimpf J, Schneider WJ. 1997. Low density lipoprotein receptor gene family members mediate yolk deposition. *The Journal of Nutrition* **127**: 801S-804S.

Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**: 1972-1973.

Carrel L, Willard HF. 2005. X-inactivation profile reveals extensive variability in X-linked gene expression in females. *Nature* **434**: 400-404.

Carroll SB. 2005. Evolution at two levels: on genes and form. *PLoS Biol* **3**: e245.

Charlesworth B. 1978. Model for evolution of Y-chromosomes and dosage compensation. *P Natl Acad Sci USA* **75**: 5618-5622.

Charlesworth B, Charlesworth D. 2000. The degeneration of Y chromosomes. *Philos Trans R Soc Lond B Biol Sci* **355**: 1563-1572.

Chickspress. 2015. Chicken Genome Consortium (ICGSC).

Chu C, Zhang QFC, da Rocha ST, Flynn RA, Bharadwaj M, Calabrese JM, Magnuson T, Heard E, Chang HY. 2015. Systematic discovery of Xist RNA binding proteins. *Cell* **161**: 404-416.

Conrad T, Akhtar A. 2011. Dosage compensation in *Drosophila melanogaster*: Epigenetic fine-tuning of chromosome-wide transcription. *Nat Rev Genet* **13**: 123-134.

Cortez D, Marin R, Toledo-Flores D, Froidevaux L, Liechti A, Waters PD, Grutzner F, Kaessmann H. 2014. Origins and functional evolution of Y chromosomes across mammals. *Nature* **508**: 488-493.

Cotton AM, Price EM, Jones MJ, Balaton BP, Kobor MS, Brown CJ. 2015. Landscape of DNA methylation on the X chromosome reflects CpG density, functional chromatin state and X-chromosome inactivation. *Hum Mol Genet* **24**: 1528-1539.

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156-2158.

Davis JK, Mittel LB, Lowman JJ, Thomas PJ, Maney DL, Martin CL, Program NCS, Thomas JW. 2011. Haplotype-based genomic sequencing of a chromosomal

polymorphism in the white-throated sparrow (*Zonotrichia albicollis*). *J Hered* **102**: 380-390.

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491-498.

Derks MFL, Schachtschneider KM, Madsen O, Schijlen E, Verhoeven KJF, van Oers K. 2016. Gene and transposable element methylation in great tit (*Parus major*) brain and blood. *BMC Genomics* **17**: 332.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**: 15-21.

Elango N, Thomas JW, Program NCS, Yi SV. 2006. Variable molecular clocks in hominoids. *P Natl Acad Sci USA* **103**: 1370-1375.

Elango N, Yi SV. 2008. DNA methylation and structural and functional bimodality of vertebrate promoters. *Mol Biol Evol* **25**: 1602-1608.

Ellegren H, Hultin-Rosenberg L, Brunström B, Dencker L, Kultima K, Scholz B. 2007. Faced with inequality: chicken do not have a general dosage compensation of sex-linked genes. *BMC Biology* **5**: 40.

Ellegren H, Smeds L, Burri R, Olason PI, Backstrom N, Kawakami T, Kunstner A, Makinen H, Nadachowska-Brzyska K, Qvarnstrom A et al. 2012. The genomic landscape of species divergence in *Ficedula* flycatchers. *Nature* **491**: 756-760.

Falls JB, Kopachena JG. 2010. White-throated sparrow (*Zonotrichia albicollis*), the Birds of North America (P. G. Rodewald, Ed.). Ithaca: Cornell lab of ornithology; Retrieved from the Birds of North America. doi:10.2173/bna.128.

Fan G, Martinowich K, Chin MH, He F, Fouse SD, Hutnick L, Hattori D, Ge W, Shen Y, Wu H et al. 2005. DNA methylation controls the timing of astrogliogenesis through regulation of JAK-STAT signaling. *Development* **132**: 3345.

Foissac S, Djebali S, Munyard K, Vialaneix N, Rau A, Muret K, Esquerre D, Zytnicki M, Derrien T, Bardou P et al. 2018. Livestock genome annotation: transcriptome and chromatin structure profiling in cattle, goat, chicken and pig. *bioRxiv*.

Gaudet F, Hodgson JG, Eden A, Jackson-Grusby L, Dausman J, Gray JW, Leonhardt H, Jaenisch R. 2003. Induction of tumors in mice by genomic hypomethylation. *Science* **300**: 489.

Grant J, Mahadevaiah SK, Khil P, Sangrithi MN, Royo H, Duckworth J, McCarrey JR, VandeBerg JL, Renfree MB, Taylor W et al. 2012. Rsx is a metatherian RNA with Xist-like properties in X-chromosome inactivation. *Nature* **487**: 254-U1511.

Graves JA. 2016. Evolution of vertebrate sex chromosomes and dosage compensation. *Nat Rev Genet* **17**: 33-46.

Hamidi T, Singh AK, Chen T. 2015. Genetic alterations of DNA methylation machinery in human diseases. *Epigenomics* **7**: 247-265.

Harris RS. 2007. Improved pairwise alignment of genomic DNA, Ph.D. Thesis. The Pennsylvania State University.

Haynes K, Killick R, Fearnhead P, Eckley I. 2016. changepoint.np: methods for nonparametric changepoint detection.

Hellman A, Chess A. 2007. Gene body-specific methylation on the active X chromosome. *Science* **315**: 1141-1143.

Hermann A, Goyal R, Jeltsch A. 2004. The DNMT1 DNA-(cytosine-C5)-methyltransferase methylates DNA processively with high preference for hemimethylated target sites. *Journal of Biological Chemistry* **279**: 48350-48359.

Horton BM, Hu Y, Martin CL, Bunke BP, Matthews BS, Moore IT, Thomas JW, Maney DL. 2013. Behavioral characterization of a white-throated sparrow homozygous for the ZAL2^{m} chromosomal rearrangement. *Behavior genetics* **43**: 60-70.

Horton BM, Hudson WH, Ortlund EA, Shirk S, Thomas JW, Young ER, Zinzow-Kramer WM, Maney DL. 2014a. Estrogen receptor alpha polymorphism in a species with alternative behavioral phenotypes. *P Natl Acad Sci USA* **111**: 1443-1448.

Horton BM, Moore IT, Maney DL. 2014b. New insights into the hormonal and behavioural correlates of polymorphism in white-throated sparrows, *Zonotrichia albicollis*. *Anim Behav* **93**: 207-219.

Horvath S. 2013. DNA methylation age of human tissues and cell types. *Genome Biol* **14**: 3156.

Howard G, Eiges R, Gaudet F, Jaenisch R, Eden A. 2007. Activation and transposition of endogenous retroviral elements in hypomethylation induced tumors in mice. *Oncogene* **27**: 404.

Huynh LY, Maney DL, Thomas JW. 2010. Contrasting population genetic patterns within the white-throated sparrow genome (*Zonotrichia albicollis*). *BMC Genet* **11**: 96.

Huynh LY, Maney DL, Thomas JW. 2011. Chromosome-wide linkage disequilibrium caused by an inversion polymorphism in the white-throated sparrow (*Zonotrichia albicollis*). *Heredity* **106**: 537-546.

Itoh Y, Kampf K, Arnold AP. 2011. Possible differences in the two Z chromosomes in male chickens and evolution of MHM sequences in Galliformes. *Chromosoma* **120**: 587-598.

Itoh Y, Melamed E, Yang X, Kampf K, Wang S, Yehya N, Van Nas A, Replogle K, Band MR, Clayton DF et al. 2007. Dosage compensation is less effective in birds than in mammals. *Journal of Biology* **6**: 2.

Itoh Y, Replogle K, Kim YH, Wade J, Clayton DF, Arnold AP. 2010. Sex bias and dosage compensation in the zebra finch versus chicken genomes: general and specialized patterns among birds. *Genome Res* **20**: 512-518.

Jarvis ED Mirarab S Aberer AJ Li B Houde P Li C Ho SY Faircloth BC Nabholz B Howard JT et al. 2014. Whole-genome analyses resolve early branches in the tree of life of modern birds. *Science* **346**: 1320-1331.

Jetz W, Thomas GH, Joy JB, Hartmann K, Mooers AO. 2012. The global diversity of birds in space and time. *Nature* **491**: 444-448.

Jetz W, Thomas GH, Joy JB, Redding DW, Hartmann K, Mooers AO. 2014. Global distribution and conservation of evolutionary distinctness in birds. *Curr Biol* **24**: 919-930.

Jukes TH, Cantor CR. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism*, pp. 21-132. Academic Press.

Julien P, Brawand D, Soumillon M, Necsulea A, Liechti A, Schütz F, Daish T, Grützner F, Kaessmann H. 2012. Mechanisms and evolutionary patterns of mammalian and avian dosage compensation. *PLoS Biol* **10**: e1001328.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772-780.

Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* **12**: 357-360.

King MC, Wilson AC. 1975. Evolution at two levels in humans and chimpanzees. *Science* **188**: 107.

Krueger F, Andrews SR. 2011. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**: 1571-1572.

Krueger F, Andrews SR. 2016. SNPsplit: allele-specific splitting of alignments between genomes with known SNP genotypes. *F1000Res* **5**: 1479.

Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* **286**: 964-967.

Laine VN, Gossmann TI, Schachtschneider KM, Garroway CJ, Madsen O, Verhoeven KJ, de Jager V, Megens HJ, Warren WC, Minx P et al. 2016. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nat Commun* **7**: 10474.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**: 357-U354.

Lechner M, Findeiss S, Steiner L, Marz M, Stadler PF, Prohaska SJ. 2011. Proteinortho: detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics* **12**: 124.

Lee I, Rasoul BA, Holub AS, Lejeune A, Enke RA, Timp W. 2017. Whole genome DNA methylation sequencing of the chicken retina, cornea and brain. *Sci Data* **4**: 170148.

Lenglos C, Calvez J, Timofeeva E. 2015. Sex-specific effects of Relaxin-3 on food intake and brain expression of corticotropin-releasing factor in rats. *Endocrinology* **156**: 523-533.

Li G, Zhang W, Baker MS, Laritsky E, Mattan-Hung N, Yu D, Kunde-Ramamoorthy G, Simerly RB, Chen R, Shen L et al. 2014a. Major epigenetic development distinguishing neuronal and non-neuronal cells occurs postnatally in the murine hypothalamus. *Hum Mol Genet* **23**: 1579-1590.

Li H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**: 3094-3100.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589-595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Proc GPD. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078-2079.

Li J, Huang Q, Zeng F, Li W, He Z, Chen W, Zhu W, Zhang B. 2014b. The prognostic value of global DNA hypomethylation in cancer: a meta-analysis. *PLoS One* **9**: e106290.

Li JX, Li RJ, Wang Y, Hu XX, Zhao YQ, Li L, Feng CG, Gu XR, Liang F, Lamont SJ et al. 2015. Genome-wide DNA methylome variation in two genetically distinct chicken lines using MethylC-seq. *BMC Bioinformatics* **16**.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**.

Lucchesi JC, Kuroda MI. 2015. Dosage compensation in *Drosophila*. *Cold Spring Harb Perspect Biol* **7**.

Maney DL. 2008. Endocrine and genomic architecture of life history trade-offs in an avian model of social behavior. *Gen Comp Endocr* **157**: 275-282.

Maney DL, Horton BM, Zinzow-Kramer WM. 2015. Estrogen receptor alpha as a mediator of life-history trade-offs. *Integr Comp Biol* **55**: 323-331.

Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**: 2867-2873.

Mank JE. 2013. Sex chromosome dosage compensation: definitely not for everyone. *Trends in Genetics* **29**: 677-683.

Mank JE, Ellegren H. 2009. All dosage compensation is local: gene-by-gene regulation of sex-biased expression on the chicken Z chromosome. *Heredity* **102**: 312-320.

Marcais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: a fast and versatile genome alignment system. *PLoS Comput Biol* **14**: e1005944.

Marin R, Cortez D, Lamanna F, Pradeepa MM, Leushkin E, Julien P, Liechti A, Halbert J, Bruning T, Mossinger K et al. 2017. Convergent origination of a *Drosophila*-like dosage compensation mechanism in a reptile lineage. *Genome Res* **27**: 1974-1987.

Martin M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* doi:10.14806/ej.17.1.200.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M et al. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**: 1297-1303.

Melamed E, Arnold AP. 2007. Regional differences in dosage compensation on the chicken Z chromosome. *Genome Biol* **8**.

Melamed E, Elashoff D, Arnold AP. 2009. Evaluating dosage compensation on the chicken Z chromosome: should effective dosage compensation eliminate sexual bias? *Heredity* **103**: 357-359.

Meyer BJ. 2010. Targeting X chromosomes for repression. *Curr Opin Genet Dev* **20**: 179-189.

Mi H, Huang X, Muruganujan A, Tang H, Mills C, Kang D, Thomas PD. 2017. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* **45**: D183-D189.

Michopoulos V, Maney DL, Morehouse CB, Thomas JW. 2007. A genotyping assay to determine plumage morph in the white-throated sparrow (*Zonotrichia albicollis*). *Auk* **124**: 1330-1335.

Mueller JC, Kuhl H, Timmermann B, Kempenaers B. 2015. Characterization of the genome and transcriptome of the blue tit *Cyanistes caeruleus*: polymorphisms, sex-biased expression and selection signals. *Molecular Ecology Resources* **16**: 549-561.

Mugal CF, Arndt PF, Holm L, Ellegren H. 2015. Evolutionary consequences of DNA methylation on the GC content in vertebrate genomes. *G3 (Bethesda)* **5**: 441-447.

Mullon C, Wright AE, Reuter M, Pomiankowski A, Mank JE. 2015. Evolution of dosage compensation under sexual selection differs between X and Z chromosomes. *Nat Commun* **6**.

Murao N, Noguchi H, Nakashima K. 2016. Epigenetic regulation of neural stem cell property from embryo to adult. *Neuroepigenetics* **5**: 1-10.

Muyle A, Zemp N, Deschamps C, Mousset S, Widmer A, Marais GA. 2012. Rapid *de novo* evolution of X chromosome dosage compensation in *Silene latifolia*, a plant with young sex chromosomes. *PLoS Biol* **10**: e1001308.

Nam K, Ellegren H. 2008. The chicken (*Gallus gallus*) Z chromosome contains at least three nonlinear evolutionary strata. *Genetics* **180**: 1131-1136.

Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Grutzner F, Kaessmann H. 2014. The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* **505**: 635-+.

Newman AM, Cooper JB. 2007. XSTREAM: a practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *BMC Bioinformatics* **8**.

Okano M, Bell DW, Haber DA, Li E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for *de novo* methylation and mammalian development. *Cell* **99**: 247-257.

Papadopulos AS, Chester M, Ridout K, Filatov DA. 2015. Rapid Y degeneration and dosage compensation in plant sex chromosomes. *P Natl Acad Sci USA* **112**: 13021-13026.

Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* **33**: 290-+.

Petkovich DA, Podolskiy DI, Lobanov AV, Lee SG, Miller RA, Gladyshev VN. 2017. Using DNA methylation profiling to evaluate biological age and longevity interventions. *Cell Metab* **25**: 954-+.

Prum RO, Berv JS, Dornburg A, Field DJ, Townsend JP, Lemmon EM, Lemmon AR. 2015. A comprehensive phylogeny of birds (Aves) using targeted next-generation DNA sequencing. *Nature* **526**: 569-U247.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841-842.

R Core Team. 2019. R: A language and environment for statistical computing. Vienna, Austria.

Ridings-Figueroa R, Stewart ER, Nesterova TB, Coker H, Pintacuda G, Godwin J, Wilson R, Haslam A, Lilley F, Ruigrok R et al. 2017. The nuclear matrix protein CIZ1 facilitates localization of Xist RNA to the inactive X-chromosome territory. *Gene Dev* **31**: 876-888.

Robertson KD. 2005. DNA methylation and human disease. *Nat Rev Genet* **6**: 597-610.

Robertson KD, A.Jones P. 2000. DNA methylation: past, present and future directions. *Carcinogenesis* **21**: 461-467.

Rodríguez-Paredes M, Esteller M. 2011. Cancer epigenetics reaches mainstream oncology. *Nature Medicine* **17**: 330.

Roeszler KN, Itman C, Sinclair AH, Smith CA. 2012. The long non-coding RNA, MHM, plays a role in chicken embryonic development, including gonadogenesis. *Dev Biol* **366**: 317-326.

Sackton TB, Grayson P, Cloutier A, Hu Z, Liu JS, Wheeler NE, Gardner PP, Clarke JA, Baker AJ, Clamp M et al. 2018. Convergent regulatory evolution and the origin of flightlessness in palaeognathous birds. *bioRxiv*.

Schmieder R, Edwards R. 2011. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27**: 863-864.

Schübeler D. 2015. Function and information content of DNA methylation. *Nature* **517**: 321.

Sharp AJ, Stathaki E, Migliavacca E, Brahmachary M, Montgomery SB, Dupre Y, Antonarakis SE. 2011. DNA methylation profiles of human active and inactive X chromosomes. *Genome Res* **21**: 1592-1600.

Shetty S, Griffin DK, Graves JA. 1999. Comparative painting reveals strong chromosome homology over 80 million years of bird evolution. *Chromosome Res* **7**: 289-295.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. *Genome Res* **19**: 1117-1123.

Smit AFA, Hubley R. 2008-2015. RepeatModeler Open-1.0.

Smit AFA, Hubley R, Green P. 2013-2015. RepeatMasker Open-4.0.

Sun D, Huh I, Zinzow-Kramer WM, Maney DL, Yi SV. 2018. Rapid regulatory evolution of a nonrecombining autosome linked to divergent behavioral phenotypes. *P Natl Acad Sci USA* **115**: 2794.

Sun D, Yi SV. 2015. Impacts of chromatin states and long-range genomic segments on aging and DNA methylation. *PLoS One* **10**: e0128517.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res* **34**: W609-612.

Suzuki MM, Bird A. 2008. DNA methylation landscapes: provocative insights from epigenomics. *Nat Rev Genet* **9**: 465-476.

Tang Y, Horikoshi M, Li W. 2016. ggfortify: unified interface to visualize statistical result of popular R packages. *The R Journal* **8**: 478-489.

Teranishi M, Shimada Y, Hori T, Nakabayashi O, Kikuchi T, Macleod T, Pym R, Sheldon B, Solovei I, Macgregor H et al. 2001. Transcripts of the MHM region on the chicken Z chromosome accumulate as non-coding RNA in the nucleus of female cells adjacent to the DMRT1 locus. *Chromosome Res* **9**: 147-165.

Thomas JW, Caceres M, Lowman JJ, Morehouse CB, Short ME, Baldwin EL, Maney DL, Martin CL. 2008. The chromosomal polymorphism linked to variation in social behavior in the white-throated sparrow (*Zonotrichia albicollis*) is a complex rearrangement and suppressor of recombination. *Genetics* **179**: 1455-1468.

Thorneycroft HB. 1966. Chromosomal polymorphism in the white-throated sparrow, *Zonotrichia albicollis* (Gmelin). *Science* **154**: 1571-1572.

Thorneycroft HB. 1975. Cytogenetic study of white-throated sparrow, *Zonotrichia albicollis* (Gmelin). *Evolution* **29**: 611-621.

Tuttle EM. 2003. Alternative reproductive strategies in the white-throated sparrow: behavioral and genetic evidence. *Behav Ecol* **14**: 425-432.

Tuttle EM, Bergland AO, Korody ML, Brewer MS, Newhouse DJ, Minx P, Stager M, Betuel A, Cheviron ZA, Warren WC et al. 2016. Divergence and functional degradation of a sex chromosome-like supergene. *Curr Biol* **26**: 344-350.

Uebbing S, Konzer A, Xu L, Backström N, Brunström B, Bergquist J, Ellegren H. 2015. Quantitative mass spectrometry reveals partial translational regulation for dosage compensation in chicken. *Molecular Biology and Evolution* **32**: 2716-2725.

Uebbing S, Künstner A, Mäkinen H, Ellegren H. 2013. Transcriptome sequencing reveals the character of incomplete dosage compensation across multiple tissues in flycatchers. *Genome Biol Evol* **5**: 1555-1566.

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* **43**: 11.10.11-11.10.33.

Wang Z, Zhang J, Yang W, An N, Zhang P, Zhang G, Zhou Q. 2014. Temporal genomic evolution of bird sex chromosomes. *BMC Evol Biol* **14**: 250.

Warren WC, Clayton DF, Ellegren H, Arnold AP, Hillier LW, Kunstner A, Searle S, White S, Vilella AJ, Fairley S et al. 2010. The genome of a songbird. *Nature* **464**: 757-762.

Weir BS, Cockerham CC. 1984. Estimating F-Statistics for the analysis of population-structure. *Evolution* **38**: 1358-1370.

Wright AE, Dean R, Zimmer F, Mank JE. 2016. How to make a sex chromosome. *Nat Commun* **7**.

Wright AE, Zimmer F, Harrison PW, Mank JE. 2015. Conservation of regional variation in sex-specific sex chromosome regulation. *Genetics* **201**: 587-+.

Wu H, Coskun V, Tao J, Xie W, Ge W, Yoshikawa K, Li E, Zhang Y, Sun YE. 2010. Dnmt3a-dependent nonpromoter DNA methylation facilitates transcription of neurogenic genes. *Science* **329**: 444-448.

Wu H, Xu T, Feng H, Chen L, Li B, Yao B, Qin Z, Jin P, Conneely KN. 2015. Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Res* **43**: e141-e141.

Wucher V, Legeai F, Hedan B, Rizk G, Lagoutte L, Leeb T, Jagannathan V, Cadieu E, David A, Lohi H et al. 2017. FEELnc: a tool for long non-coding RNA annotation and its application to the dog transcriptome. *Nucleic Acids Res* **45**: e57.

Yang ZH. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* **24**: 1586-1591.

Yi SJ, Charlesworth B. 2000. Contrasting patterns of molecular evolution of the genes on the new and old sex chromosomes of *Drosophila miranda*. *Molecular Biology and Evolution* **17**: 703-717.

Zelic R, Fiano V, Zugna D, Grasso C, Delsedime L, Daniele L, Galliano D, Pettersson A, Gillio-Tos A, Merletti F et al. 2016. Global hypomethylation (LINE-1) and gene-specific hypermethylation (GSTP1) on initial negative prostate biopsy as markers of prostate cancer on a rebiopsy. *Clin Cancer Res* **22**: 984.

Zeng J, Konopka G, Hunt BG, Preuss TM, Geschwind D, Yi SV. 2012. Divergent whole-genome methylation maps of human and chimpanzee brains reveal epigenetic basis of human regulatory evolution. *Am J Hum Genet* **91**: 455-465.

Zhang J. 2000. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *Journal of Molecular Evolution* **50**: 56-68.

Zhang M, Yan FB, Li F, Jiang KR, Li DH, Han RL, Li ZJ, Jiang RR, Liu XJ, Kang XT et al. 2017. Genome-wide DNA methylation profiles reveal novel candidate genes associated with meat quality at different age stages in hens. *Sci Rep-Uk* **7**.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol* **9**: R137.

Zhou Q, Bachtrog D. 2012. Chromosome-wide gene silencing initiates Y degeneration in *Drosophila*. *Curr Biol* **22**: 522-525.

Zhou Q, Zhang JL, Bachtrog D, An N, Huang QF, Jarvis ED, Gilbert MTP, Zhang GJ. 2014. Complex evolutionary trajectories of sex chromosomes across bird taxa. *Science* **346**: 1332-+.

Zimmer F, Harrison PW, Dessimoz C, Mank JE. 2016. Compensation of dosage-sensitive genes on the chicken Z chromosome. *Genome Biol Evol* **8**: 1233-1242.

Zinzow-Kramer WM, Horton BM, McKee CD, Michaud JM, Tharp GK, Thomas JW, Tuttle EM, Yi S, Maney DL. 2015. Genes located in a chromosomal inversion are correlated with territorial song in white-throated sparrows. *Genes Brain Behav* **14**: 641-654.